# Utilizing Machine Learning for the Early Detection of Coronary Heart Disease

**Mudhafar Jalil Jassim Ghrabat**
Iraqi Commission for Computers and Informatics, The Informatics Institute for Postgraduate Studies, Baghdad 10013, Iraq | Design and IoT Lab, Al-Turath University, Baghdad, 10013, Iraq
mudhafar.jalil@iips.icci.edu.iq

**Siamand Hassan Mohialdin**
Nursing College, Hawler Medical University, Erbil, Iraq
syamand.mohialdin@hmu.edu.krd

**Luqman Qader Abdulrahman**
College of Health Science, Hawler Medical University, Erbil, Iraq
luqman.qader@hmu.edu.krd

**Murthad Hussein Al-Yoonus**
Department of Information Technology, Noble Technical Institute, Erbil, Iraq
marthad.hussain@noble.edu.krd

**Zaid Ameen Abduljabbar**
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, 61004, Iraq | Huazhong University of Science and Technology, Shenzhen Institute, Shenzhen, 518000, China
zaid.ameen@uobasrah.edu.iq (corresponding author)

**Dhafer G. Honi**
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, 61004, Iraq | Department of IT, University of Debrecen, Debrecen, 4002, Hungary
dhafer.honi@uobasrah.edu.iq

**Vincent Omollo Nyangaresi**
Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science & Technology, Bondo, 40601, Kenya | Department of Applied Electronics, Saveetha School of Engineering, SIMATS, Chennai, Tami lnadu, 602105, India
vnyangaresi@jooust.ac.ke

**Iman Qays Abduljaleel**
Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, 61004, Iraq
iman.abduljaleel@uobasrah.edu.iq

**Husam A. Neamah**
Department of Electrical Engineering and Mechatronics, Faculty of Engineering, University of Debrecen, Debrecen, 4028, Ótemető u.4-5, Hungary
husam@eng.unideb.hu (corresponding author)

*Received: 18 June 2024 | Revised: 15 July 2024 | Accepted: 20 July 2024*

## ABSTRACT

**Coronary Heart Disease (CHD) is a persistent health issue, and risk prognosis is very important because it creates opportunities for doctors to provide early solutions. Despite such promising results, this type of analysis runs into several problems, such as accurately handling high-dimensional data because of the abundance of extracted information that hampers the prediction process. This paper presents a new approach that integrates Principal Component Analysis (PCA) and feature selection techniques to improve the prediction performance of CHD models, especially in light of dimensionality consideration. Feature selection is identified as one of the contributors to enhance model performance. Reducing the input space and identifying important attributes related to heart disease offers a refined approach to CHD prediction. Then four classifiers were used, namely PCA, Random Forest (RF), Decision Trees (DT), and AdaBoost, and an accuracy of approximately 96% was achieved, which is quite satisfactory. The experimentations demonstrated the effectiveness of this approach, as the proposed model was more effective than the other traditional models including the RF and LR in aspects of precision, recall, and AUC values. This study proposes an approach to reduce data dimensionality and select important features, leading to improved CHD prediction and patient outcomes.**

***Keywords-random forest; decision trees; Principal Component Analysis (PCA); Machine Learning (ML) classifiers; Coronary Heart Disease (CHD); hyperparameters; prediction***

## I. INTRODUCTION

Cardiovascular diseases are the leading cause of mortality worldwide, responsible for around 18 million deaths per year. Although there have been notable advances in medical technology, accurately predicting cardiac disease remains a challenging task. Existing prediction approaches often demonstrate subpar performance, resulting in missed diagnoses and postponed actions. Coronary Heart Disease (CHD), often known as Cardiovascular Disease (CVD), is the leading cause of death worldwide. As a result, several studies have been conducted on the early detection of cardiac problems and the identification of the most important associated risk factors. Despite extensive efforts, prediction accuracy has remained inadequate, and the identification of the most influential risk factors has been challenging [1]. Data analysis approaches have been used to help healthcare professionals detect early indications of cardiac disease. Prospective patients may undergo many diagnostic tests to mitigate the burden of experiencing such a condition [2-3]. Accurate methods for predicting the early stages of heart disease can be crucial to preserving people's lives.

Deep Learning (DL) is transforming various applications and has attracted worldwide attention for its disruptive influence across multiple domains. Image analysis and predictive modeling can improve medical evaluation and treatment planning in healthcare. The automotive industry is utilizing DL for self-driving systems to make transport safer and more efficient. DL is enhancing fraud detection and trading strategies in finance. Retailers use DL to make personalized recommendations and manage inventories. Furthermore, DL is changing natural language processing, powering virtual assistants and chatbots in the customer service and entertainment industries. ML-based methods usually include single or ensemble classification techniques, using feature extraction/selection methods to improve results [4-5]. DL algorithms are widely applied to detect CHD. Most approaches divide an experimentation dataset into two portions for training and testing and then employ classification techniques to create prediction models from training samples. Consequently, models are much more likely to be fitted to consistently distributed datasets and mislabel unevenly distributed (biased) data [6]. In addition, common ML classifiers, such as Adaboost, Gradient Boosting (GB), Extra Trees (ET), and CatBoost, are employed with tried-and-true methods, such as Principal Component Analysis (PCA), for feature selection. The number of DL applications has recently increased, showing its ability and adaptability, exemplified by models such as DCNNBT, which is a classifier for brain tumors based on deep convolutional neural networks. Deep neural networks generate more precise and detailed conclusions about images compared to humans, using approaches that depend solely on subtle patterns detected by convolutional filters. In addition, data augmentation along with transfer learning make the models more flexible, improving the detection of brain tumors. As these approaches are fairly easy to scale and have good fault tolerance, they are well-suited for large-scale use. A rather practical improvement based on U-Net-style models in image segmentation is a leap forward toward a better definition of pathological and structural margins. DL applications also help to improve the capacity of clinicians to make sound judgments about treatment and diagnosis. The consequences extend to other areas, establishing new methods for processing the results of biological experiments. For example, beyond cancer diagnosis, DL can be used to predict the efficacy of anti-angiogenic drugs. These milestones illustrate the vast scope and profound influence of DL in handling wide-ranging and complex issues in several fields, highlighting its disruptive potential for innovation in solving practical concerns [7].

PCA is a statistical method that employs an orthogonal transformation to turn a collection of observations of potentially correlated variables into sets of linearly uncorrelated values, known as the principal components, which are the directions in the data that have the most variation. This study selected five essential factors to pre-diagnose Coronary Artery Disease (CAD): Age, hypertension, usual chest pain, T-wave inversion, and localized wall motion anomalies. Then, different search techniques were used to select relevant features. Afterward, PCA and the AdaBoost algorithm were employed for the classification task, achieving the most exceptional performance in a publicly available dataset with such a limited number of features. This strategy achieved enhanced efficiency, precision, and resilience compared to previous methods that relied on a wider array of features. This distinctive approach makes it a viable instrument for the early identification of CHD, as it specifically addresses certain obstacles such as data integration, algorithmic bias, and clinical validation [8]. Previous studies used a range of 16 to 40 features, whereas the proposed technique requires only five. By reducing the number of features, the model becomes less complicated and more efficient, while maintaining accuracy. The proposed method incorporates eight feature selection search techniques to ensure that the selected features align with the desired criteria. Preprocessing effectively prepares the data for the subsequent stage. PCA reduces data dimensionality by transforming it into a new space and focusing on the most important aspects. This not only decreases computational cost but also improves its results as the amount of useless data decreases.

Accurate determination of CHD remains a challenge. Therefore, this study presents an improved approach to improve prediction accuracy using PCA and feature selection. This approach provides a practical solution to manage circadian timing disturbances and improve the prediction of CHD risk. By applying the proposed method, the number of features is reduced from the 13 features of the dataset, and PCA overlays the principal components for better predictions. This also helps to reduce computational complexity and improve the generalizability and validity of classifiers from irrelevant and unnecessary attributes. Improvements to current prediction models need to focus on the parameters used for PCA analysis and try other methods of dimension reduction. These adjustments will assist in building a model that will offer desirable results in certain healthcare segments. In addition, this study compared four different ML classifiers for CHD diagnosis to determine which is the most appropriate. As such, it should be noted that new insights and trends have been discovered, such as ranking features, human-interpretable models, and new correlations, for future progress in CHD prediction. When coupled with prompt human interventions and customized treatment procedures, the proposed approach has the potential to not only save lives but also save healthcare costs.

The integration of a dimensionality reduction and several categorization models aimed to achieve two primary goals: (i) effectively represent the characteristics of the dataset, and (ii) use ML methods to produce accurate predictions. Various classification models, including GB, AdaBoost, Catboost, and ET, were employed with specific hyperparameter values. The primary contributions of this study are as follows:

- Examines several ML classifiers and classification techniques to predict and detect heart disease.
- Preprocesses the data to solve the incompleteness and unreliability problems of heart disease datasets.
- Uses PCA to select features by extracting the most valuable, enhancing classification results by reducing dimensionality.
- Provides increased classification and prediction results by implementing and comparing different ML classifiers.
- Compares advanced ML classifiers with existing classification methods, according to performance measures, to predict whether a patient has heart disease.

## II. LITERATURE REVIEW

Numerous studies have been conducted on predicting CHD, employing different types of ML techniques using various feature selection approaches [9, 10]. However, the recognition of many features is more difficult, while large feature dimensions increase training times. In [11], a prediction model for CHD was proposed using supervised ML methods, namely Bernoulli Naive Bayes (BNB), Random Forest (RF), and Gaussian Naive Bayes (GNB). In addition, several related risk factors were discussed. A large dataset was used from the Cleveland database of the UCI repository. The results showed that the accuracy of the GNB and BNB models was 85%, while the accuracy of the RF model was 75%. Furthermore, the accuracy, recall, and F1 scores of GNB and BNB were higher compared to those of RF, which was attributed to RF randomization. In [12, 13], DL techniques were used to predict heart disease. In [14], an improved 3D U-net CNN model was proposed to segment the coronary artery and predict the risk of heart attack. This model was implemented using various datasets and two contexts: one without and another with a centerline. A Deep Belief Network (DBN) was used to analyze ventricular data, which differentiated between left and right ventricles with a dice coefficient of 0.8291 and an overall accuracy of 78%. Despite improved results, such as those in the Fully Convolutional Network (FCN) applied for image segmentation, the intricacies of medical images were a concern and showed that, while deep learning has promise for medical imaging, it is not without issues.

In [15, 16], ML and DL techniques were combined to predict the risk of CHD. In [16], two types of experiments were carried out to predict the risk of CHD, trained on a logically consistent dataset acquired from the Korean National Health and Nutritional Examination Survey. PCA and variational autoencoder increased the overall performance of a single DNN, while comparing the proposed and other ML methods indicated that the former was superior to the latter, as it achieved higher accuracy (0.892), specificity (0.840), precision (0.911), recall (0.920), f-measure (0.915), and AUC (0.882). In [17], various ML and DL techniques were applied and compared in the UCI HD dataset. The dataset contains certain unimportant attributes that were removed with an isolation forest to enhance the validity of the findings. In addition, the data were normalized and an accuracy of 94.2% was achieved with the proposed DL model. However, the sample size of the dataset was small. In [18], nine typical ML and DL classifiers were compared to predict CHD, with LR achieving the best performance. In [19], RF outperformed the other models, as it was highly sensitive to detect the target (98%) but had low specificity (54%). In [20], RF, LR, and SVM were compared in CHD detection, with RF achieving 92.9% accuracy. This study employed an extensive dataset and models were trained with cross-validation to reduce overfitting. In [21], iris images were used to diagnose CHD, using wavelet transform, GLCM, and GLRLM to extract features from preprocessed images. SVM achieved 93% accuracy. Telemedicine applications may benefit from such methods [22].

However, ML approaches face many problems in CHD prediction and detection. The segmentation process begins with the identification and positioning of the coronary arteries. Manually identifying coronary arteries is time-consuming, and biases can impact segmentation results. In addition, medical images have certain qualities. At the first point, it is hard to tell whether a medical image is healthy or unhealthy since the structure of the two images is so similar. From the above review of the existing approaches, certain major issues in heart disease detection can be observed:

- Predictive models cannot learn effectively from real-world heart disease datasets in case of imbalanced data and variations.
- Accuracy is insufficient when a model is tested on real-world data that may differ significantly from the dataset it was trained. The dataset must be normalized to prevent the model from overfitting.
- Most currently accessible prediction models were trained to utilize the entire training set or a randomly selected subset.
- Several approaches have been implemented on the Cleveland dataset, although their accuracies are highly dependent on time calculations.
- No strategy has been proposed to generate training data by discriminating between ordinary and extremely biased subgroups to create a reliable prediction model.
- A small dataset can cause DL models to perform unsatisfactorily.

TABLE I. COMPARISON OF EXISTING AND PROPOSED APPROACHES FOR HEART DISEASE DETECTION

| Problem | Existing approaches | Proposed approach |
|---|---|---|
| Imbalanced datasets | Traditional models struggle to learn from imbalanced datasets | The proposed method utilizes a hybrid strategy that integrates oversampling, undersampling, and synthetic minority oversampling strategies to tackle an imbalanced dataset. |
| Overfitting and generalizability | Models trained on entire training sets or random subsets may overfit and lack generalizability to real-world data | The proposed method utilizes k-fold cross-validation to avoid overfitting and enhance generalizability |
| Feature selection | Traditional feature selection methods may fail to capture subtle variations and important features | The proposed method employs a fusion of PCA and correlation analysis to detect the most significant and useful features |
| Sample size limitations | Small sample sizes can negatively affect the performance of DL algorithms | The proposed method employs a transfer learning approach to leverage pre-trained DL models and improve performance with limited data |

The segmentation performance of CHD detection can improve with improving dataset quality. Identification of several features is harder, leading to a lack of features and hardly noticeable variations in practice. In addition, several potential causes affect medical data collection, including but not limited to variations in patients, equipment, variables, as well as operating environments. Table I summarizes the key differences between the proposed and existing approaches to CHD detection, highlighting the challenges faced by the existing approaches and how the proposed method addresses them.

## III. PROPOSED METHOD

### A. *Statement of the Problem*

There is a research gap in early coronary prediction. Although many studies focused on predicting the risk of CHD, there seems to be a deficiency in early prediction using clinical criteria [12]. Regarding early CHD, due to plaque buildup that results in a reduced blood supply to the cardiac muscle, there is still scarce evidence. Although techniques such as angiography show efficiency in detecting narrowed arteries, non-invasive AI-based algorithms can have more accurate performance [20]. This study presents a novel computational model to fine-tune early CHD diagnosis, investigates several classifiers in CHD, and validates their effectiveness [23].

### B. *Proposed Method*

One of the primary causes of heart disease is a variety of conditions affecting the heart, including blood vessel diseases, arrhythmias, etc. This study aimed to develop a model to predict CHD. The data were preprocessed to clean them, including checking for missing values, standard scaling, etc. Then, essential features were selected using PCA to reduce the data dimensions and speed up model training. Then, data were split into training and testing sets. Finally, classification was performed using four ML classifiers, namely GB, AdaBoost, CatBoost, and ET with their respective hyperparameter settings, to predict heart diseases. The performance of the classifiers was evaluated using F1-score, accuracy, precision, and recall metrics. Figure 1 shows a block diagram of the proposed method to CHD.

#### 1) *Dataset*

The Heart Disease Dataset [24] is a summarized version of the UCI Heart Disease dataset [25], comprising 14 attributes, most of them integer and float datatypes. The UCI Heart Disease dataset [25] comprises four databases, namely Long Beach V, Cleveland, Hungary, and Switzerland, with 76 attributes. However, the dataset in [24] focuses on fourteen key characteristics and Cleveland. The primary target attribute indicates the presence of heart disease (0 for absence of disease, 1 for disease). Age, sex, ECG results, blood pressure, fasting blood sugar, chest pain type, cholesterol levels, exercise-induced angina, ST segment slope, major vessel count, maximum heart rate, ST depression, and thallium test results provide information on the patient's health and heart condition. The dataset is useful for exploring the correlation between these characteristics and CHD while protecting patients' privacy.

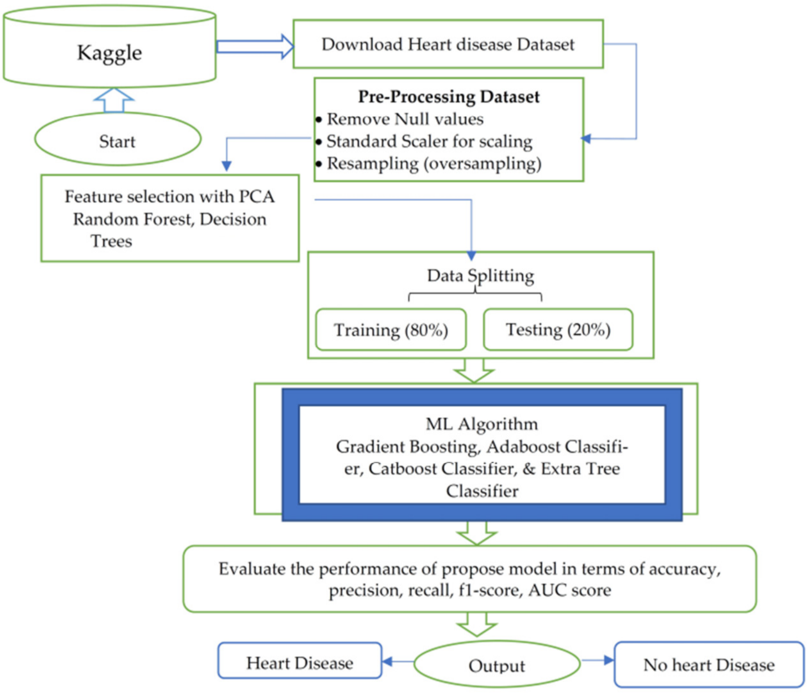


Fig. 1. Block diagram of the proposed CHD prediction model.

### 2) *Data Preparation/Preprocessing*

Preprocessing is used to get data that is comprehensive, coherent, and easily understandable. The results of ML algorithms are influenced by data quality. High-quality data leads to informed and accurate decision-making. Data were checked, null values were checked, and a standard scaler was used to standardize the input features. The standard score of each feature was measured using

$$y = (x - mean) \,/\, SD \tag{1}$$

Data cleaning and normalization were performed to ensure uniformity and efficiency during training and testing. Data normalization is essential to achieve the best possible performance with ML algorithms since it standardizes the data and brings all features to a consistent scale. Normalization results in improved model convergence, larger-scale features are not allowed to dominate, and training is more stable and effective as a whole.

### 3) *Feature Selection with PCA, DT, and RF*

Feature selection is an essential and critical stage in improving CHD prediction models. PCA enables the transformation of data into a new coordinate system while retaining important information. This, in turn, simplifies the analysis process. Although PCA might serve as an alternative option, there are some challenges. While dimensionality reduction can help in the initial detection of issues, it can eliminate data and weaken a model's prediction ability. However, since PCA implies that all connections are linear, it may lead to a simplified depiction of the nonlinearities that are evident in several predictors of CHD. Despite the aforementioned points, it remains uncertain to what extent PCA contributes to the prediction of cardiac events and the balance between the elimination of irrelevant variables without compromising the potential predictive ability of models [26].

RF is a supervised ML algorithm that falls under the broader category of ensemble learning. As indicated by its name, it is a forest composed of randomly constructed Decision Trees (DTs). RF is based on the bagging approach. It involves running the same process numerous times and calculating the error for each run. These errors are then aggregated to provide a more accurate and improved result. Bagging involves the creation of many DTs that are produced individually and then merged to provide a final result. The algorithm randomly selects features and then divides the node based on the feature that will be responsible for constructing particular splits in the creation process. This effect is further enhanced by including random breakpoints in each feature. RF analyzes the data to assign scores to each characteristic. The significance of a feature is measured by calculating the total impurity reduction across all nodes that make decisions based on this information. Furthermore, RF exhibits resistance to high energy levels.

The DT approach is very adaptable and widely accepted because of its simplicity, which is considered one of its key strengths. The system consists of three nodes: the chance node, the choice node, and the end node. The chance node represents the expected result of a certain function, whereas the decision node indicates the possibility of two outcomes after invoking the service until another call is made. The end node is the ultimate node that concludes every trip with a conclusive outcome. A DT starts with a root node, which then divides into several branches or nodes. Each node contains data-specific information on the optimal choice to make at a particular testing stage, while each link represents a decision rule. There are two methods for displaying the tree: using the Gini index and using the entropy rule as a criterion. DT is one of the most straightforward and transparent forecasting models.

PCA has been used in conjunction with RF and DT for early detection of CHD in many studies. PCA has the following characteristics:

- Combining PCA with RF and DT can utilize the recorded data more effectively. This approach is particularly useful for identifying early CHD cases.
- PCA decreases the number of dimensions in a dataset and converts it into a new space while retaining crucial characteristics. This reduces computational load and enhances the performance of RF and DT. This is achieved by picking the best features that provide the best explanatory power for the nodes. This combination yields significantly improved results.
- PCA can possibly mitigate the issue of multicollinearity by generating uncorrelated principal components.
- PCA is preferred to ensure that models such as RF or DT are in a stable state, as they tend to fail when dealing with correlated information.
- PCA can effectively identify the most significant features, making it the optimal approach for feature selection and leading to more accurate predictions.
- Combining PCA with AdaBoost and RF can increase accuracy, particularly when using a limited number of features. Efficiency is crucial since it allows for quick and accurate diagnosis, facilitating timely treatment.

Applying PCA in combination with RF and DT can effectively identify CHD at an early stage, serving as a significant approach for preventive ML methods.

#### 4) *Classification and Prediction*

Following PCA-based feature selection, the dataset was partitioned into a training and a testing set with an 80:20 ratio. This proportion is meant to provide a medium between providing the model with enough information to learn patterns and a fair assessment of its generalizability. Four ML classifiers were used to detect CHD in a patient:

- AdaBoost [27, 28]
- Gradient Boosting (GB) [29]
- Extra Tree (ET) [30]
- CatBoost [31]

The choice of classifiers is essential for achieving accurate results. In situations when both predicted accuracy and interpretability are crucial, these classifiers were chosen due to their distinct advantages in handling classification tasks.

### C. *Novelty of the Method*

This study employed RF, DT, and PCA for feature reduction and commonly used ML classifiers, such as Adaboost, GB, ET, and CatBoost. At first, these techniques are applied in the specific context of CHD prediction, demonstrating domain-specific expertise in addressing the unique challenges associated with cardiovascular health data. The application of RF, DT, and PCA for CHD prediction was optimized through hyperparameter fine-tuning and exploring relevant dimensionality reduction techniques. These customizations ensure the model's effectiveness in the specific domain. Selecting the four ML classifiers is driven by carefully considering their suitability for CHD prediction. Furthermore, the study uncovers unique insights and patterns in CHD prediction that may not have been explored in existing literature, including specific feature importance rankings, model interpretability, and unexpected correlations, providing novel contributions to the field. Lastly, this study underscores the practical significance of accurate CHD prediction, which can lead to early interventions and personalized treatment plans, potentially saving lives and reducing healthcare costs.

### D. *ML Classifiers*

The ML models were used individually. This approach allows for assessing the strengths and weaknesses of each model in the context of CHD prediction and making informed decisions on model selection and hyperparameter tuning.

#### 1) *AdaBoost (Adaptive Boosting)*

Adaboost is a method to increase the performance of weak learners, such as DT, by giving weights to data points and iteratively changing them to focus on misclassified cases. Adaboost was chosen due to its ability to improve the accuracy of weak classifiers, making it suitable for improving the predictive power of basic models. Adaboost can be used alone or in conjunction with other methods to improve their performance. Algorithm 1 describes how AdaBoost works.

```
Algorithm 1: AdaBoost classifier
Input: Heart disease dataset
Output: Classification results

1: At the start, AdaBoost selects a training sample through the random selection process.
2: AdaBoost is trained iteratively by selecting a training dataset depending on the accuracy of such a prediction using the most recent training.
3: It gives the observations that were incorrectly assigned a greater weight to ensure that, in the subsequent iteration, those observations would be assigned a high likelihood of classification.
4: Determines how much weight to give the trained classifier for each iteration by evaluating its classification accuracy. A classifier that offers more accurate results will be assigned more weight.
5: This method continues to execute until the entire training data is fitted with no errors or the maximum number of set estimators has been met.
```

```
6: Vote across all the classifiers.
```

Boosting is an ensemble method that employs multiple weak learners to improve a single set of data. Classification performance can be enhanced through iterative boosting. AdaBoost is the most widely used boosting technique and has been extensively investigated because it was the first practical boosting algorithm. AdaBoost involves a user-determined number of iterations. For each cycle, the complete training set is fed into a trio of weak learners, and the results are compared to the expected labels of samples. The resulting error function is used to assign weight to each spectrum in the subsequent iteration. This results in misclassified data given more weight and correctly classified ones given less. Data that have not been properly classified will be the subject of subsequent iterations. Every iteration's weighted weak learners are given access to the unseen data to establish the projected class of a previously unseen test subject. The winner is then decided by weighted majority voting [27]. Algorithm 2 describes the proposed model using the AdaBoost classifier.

```
Algorithm2. Proposed prediction model
Input: Heart disease Cleveland data

1: Heart disease data collection
2: Prepare the data with a preprocessing
mechanism
3: Null value check
4: Standard scaler
5: Dimensionality reduction using RF, DT,
and PCA. Select essential features
6: Dataset separation into Training (80%)
and Testing (20%)
7: Train four machine learning
classifiers with hyperparameter tuning:
AdaBoost, GB, ET, and CatBoost
8: Calculate classification results in
terms of performance metrics
9: Test the models
10: Provide prediction results using
confusion matrices
11: Output: Heart disease prediction (Yes
or No)
```

### 2) *Gradient Boosting (GB)*

GB builds an ensemble of DTs, with each tree correcting the flaws of the previous. GB is well known for its great prediction accuracy and ability to handle complex data interactions. It can be employed alone or as an ensemble, but it is frequently used as an independent model due to its high predictive ability.

### 3) *Extra Trees (ET)*

ET is an ensemble learning technique based on DT, commonly called extremely randomized trees, that incorporates randomness into tree construction to minimize overfitting. ET is used because of its resistance to overfitting and ability to capture significant features while lowering variation. ET can be used individually or as an ensemble. Some studies use it as an individual model to benefit from its noise-reduction properties.

### 4) *CatBoost*

CatBoost is an effective gradient-boosting method for working with categorical features. CatBoost is selected due to its ability to perform well with a mixture of categorical and numeric data, commonly encountered in healthcare datasets. CatBoost is usually used independently, benefiting from its category feature processing.

## IV. EXPERIMENTAL ENVIRONMENT

The models were developed on an HP workstation with Windows 10, 1TB HDD, i7 CPU, and 32GB of RAM, using Python and Jupyter Notebook.

### A. *Feature Details*

Table II shows the subset of features used [24].

TABLE II. FEATURE DETAILS OF DATASET.

| Features | Value |
|---|---|
| AGE | Age (in years) |
| SEX | 1: male, 0: female |
| Chest Pain (CP) type | 1: typical angina, 2: atypical angina, 3: non-angina pain, 4: asymptomatic |
| Treetops | Systolic blood pressure at rest (in mmHg on admission to hospital) |
| Chol | Serum cholesterol in mg/dl |
| FBS | Fasting blood sugar > 120 mg/dl (1: true, or 0: false) |
| Restecg (Resting electrocardiographic results) | 0: normal, 1: having ST-T wave abnormality (T-wave inversions and ST elevation or depression > 0.05 mV), 2: showing probable or definite left ventricular hypertrophy by Estes' criteria |
| Exam | Exercise-induced angina (1: yes. 0: no) |
| Halacha | Maximum heart rate achieved |
| Slope (slope of peak exercise ST segment) | 1: upsloping, 2: flat, 3: downsloping |
| Old peak | Exercise-induced ST depression relative to rest |
| Target | 1: Yes, 0: No |
| Thal | Exercise Thallium heart scan:<br>3: normal, 6: fixed defect; 7: reversible defect |
| Ca | No. of major vessels (0 to 3) colored by fluoroscopy |

### B. *Hyperparameter Settings*

The following hyperparameters were chosen through experimentation and optimization to achieve the best possible performance for each ML classifier in the given context.

#### 1) *AdaBoost*

- random_state: Setting this parameter to 0 ensures that the random number begins from the same initial state each time the model is trained, ensuring reproducibility of results.
- n_estimators: AdaBoost is a method for building robust ensemble models by combining multiple weak learners (typically DTs). This ensemble employs 98 DTs.

#### 2) *Gradient Boosting (GB)*

- n_estimators: GB also creates an ensemble by combining several DTs. In this case, 100 DTs were used.
- max_depth: Each DT's maximum depth is set to one option, making it shallow. This prevents overfitting and ensures that each tree is a simple learner.

- learning_rate: A learning rate of 1.0 indicates that the contribution of each tree to the ensemble is not scaled down, resulting in aggressive learning.

### *3) Extra Trees (ET)*

- n_estimators: ET uses several randomized DTs. In this case, the ensemble was built using 6 DTs.
- random_state: Setting it to 98, as in AdaBoost, ensures reproducibility by seeding the random number with a specific seed.

### *4) CatBoost*

- Iterations: Like AdaBoost and GB, CatBoost builds an ensemble from independent decision trees. In this study, 5 iterations were performed.
- random_seed: This parameter ensures that results can be reproduced by providing a seed for the random number generator.
- learning_rate: CatBoost uses a 0.6 learning rate to regulate the gradient-boosting step size. Increasing the learning rate can hasten convergence, but fine-tuning may be necessary.

## *C. Performance Measurement Parameters*

Performance metrics are essential to evaluate, validate, and compare the proposed system. However, there are many performance parameters available for classification and error measurement. For this purpose, classification performance parameters and confusion matrix are used.

### *1) Confusion Matrix (CM)*

CM is a table-based representation of ground-truth labeling versus predicted results. Each row in CM signifies the cases in the predicted class, whereas every column represents cases inside an actual class. It serves as a foundation to calculate metrics to analyze the results.

TABLE III. CONFUSION MATRIX

| | | **Actual** | |
|---|---|---|---|
| | | **Heart attack (Yes)** | **Heart attack (No)** |
| **Predicted** | **Heart attack (Yes)** | TP | FP |
| | **Heart attack (No)** | FN | TN |

TP (True Positive) denotes how many positive class instances the model predicts properly, FP (False Positive) represents how many negative class instances the model predicts wrongly, FN (False Negative) represents the number of positive class instances that the model predicts wrongly, and TN represents the number of negative instances that the model predicts correctly [7, 29].

## *D. Classification Metrics*

Classification accuracy is determined as the number of true predictions divided by the number of total predictions [32]. Precision is the ratio of TPs to overall expected positives [6]. Recall is the ratio of TP across all positives in the ground truth. F1-score is the harmonic mean between recall and precision. AUC-ROC score uses TPR and FPR. The corresponding equations are:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{4}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{6}$$

TPR, or recall, represents a portion of positive samples correctly labeled as positive compared to all positive samples. FPR, or fallout, is the ratio of the negative sample erroneously interpreted positively compared to all negative samples. Alternatively, the more negative samples mistakenly classify as positive, the higher the FPR. The Precision-Recall curve depicts trade-offs across recall and precision for various threshold values. A high AUC suggests strong recall and high precision [6].

# V. RESULTS

## *A. Exploratory Data Analysis Visualization*

Different visualization representations display several features in a dataset. This study considered various graphics visuals to examine the data and handle issues [33]. Figure 2 shows a heatmap plot of the dataset to display the null values, which present null values for each feature [31, 34].

Figure 4 displays the age distribution of a group of individuals, ranging from 44 to 62 years old. The counts for each age category vary, ranging from 0.5 for age 44 to 17 for age 62. Figure 5 shows the proportion of male and female patients in the dataset. Figure 3 displays a correlation matrix between each feature. Its cell value is shown from minus one to plus one through different color levels. Brown indicates the highest correlation and dark blue indicates the lowest one. Figure 6 shows three box plots that compare the age distribution for different categories. The categories are sex, cp, and slope, shown on a horizontal axis. The age is shown on a vertical axis. The box plots are colored orange and blue, indicating each category's output values of 0 and 1, and show the median, quartiles, and range of age for each category and output value. The outliers are marked as black dots. Figure 7 shows the relationships between age, thalach, sex, and output. The graph on the left shows how thalach varies with age for different sexes. The graph on the right shows how thalach varies with age. The lines are colored orange and blue, indicating sex 0 and 1 on the left graph and output 0 and 1. Figure 8 shows a set of histograms and density plots that visualize the distribution of different variables in the dataset. The image has 14 plots, each for each variable. The x-axis of each plot shows a range of values for the variable, and the y-axis shows a count of observations. The plots also show the density curve of the variable, which indicates how likely a value is to occur.

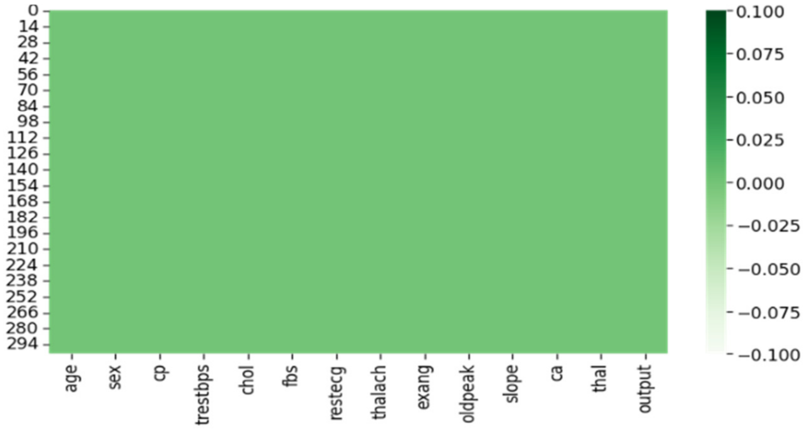
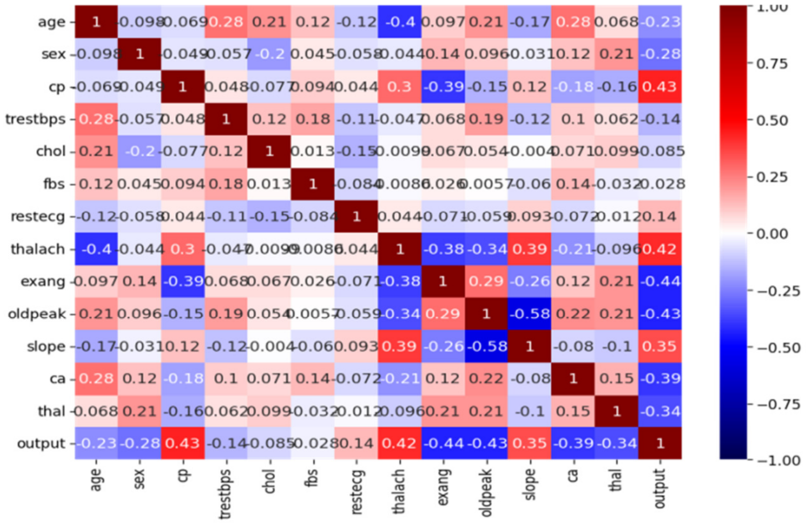


Fig. 2. Heatmap for null values.



| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1 | -0.098 | -0.069 | 0.28 | 0.21 | 0.12 | -0.12 | -0.4 | 0.097 | 0.21 | -0.17 | 0.28 | 0.068 | -0.23 |
| sex | -0.098 | 1 | -0.049 | -0.057 | -0.2 | 0.045 | -0.058 | -0.044 | 0.14 | 0.096 | -0.031 | 0.12 | 0.21 | -0.28 |
| cp | -0.069 | -0.049 | 1 | 0.048 | -0.077 | 0.094 | 0.044 | 0.3 | -0.39 | -0.15 | 0.12 | -0.18 | -0.16 | 0.43 |
| trestbps | 0.28 | -0.057 | 0.048 | 1 | 0.12 | 0.18 | -0.11 | -0.047 | 0.068 | 0.19 | -0.12 | 0.1 | 0.062 | -0.14 |
| chol | 0.21 | -0.2 | -0.077 | 0.12 | 1 | 0.013 | -0.15 | -0.0099 | 0.067 | 0.054 | -0.004 | 0.071 | 0.099 | -0.085 |
| fbs | 0.12 | 0.045 | 0.094 | 0.18 | 0.013 | 1 | -0.084 | -0.0086 | 0.026 | 0.0057 | -0.06 | 0.14 | -0.032 | -0.028 |
| restecg | -0.12 | -0.058 | 0.044 | -0.11 | -0.15 | -0.084 | 1 | 0.044 | -0.071 | -0.059 | 0.093 | -0.072 | -0.012 | 0.14 |
| thalach | -0.4 | -0.044 | 0.3 | -0.047 | -0.0099 | -0.0086 | 0.044 | 1 | -0.38 | -0.34 | 0.39 | -0.21 | -0.096 | 0.42 |
| exang | 0.097 | 0.14 | -0.39 | 0.068 | 0.067 | 0.026 | -0.071 | -0.38 | 1 | 0.29 | -0.26 | 0.12 | 0.21 | -0.44 |
| oldpeak | 0.21 | 0.096 | -0.15 | 0.19 | 0.054 | 0.0057 | -0.059 | -0.34 | 0.29 | 1 | -0.58 | 0.22 | 0.21 | -0.43 |
| slope | -0.17 | -0.031 | 0.12 | -0.12 | -0.004 | -0.06 | 0.093 | 0.39 | -0.26 | -0.58 | 1 | -0.08 | -0.1 | 0.35 |
| ca | 0.28 | 0.12 | -0.18 | 0.1 | 0.071 | 0.14 | -0.072 | -0.21 | 0.12 | 0.22 | -0.08 | 1 | 0.15 | -0.39 |
| thal | 0.068 | 0.21 | -0.16 | 0.062 | 0.099 | -0.032 | -0.012 | -0.096 | 0.21 | 0.21 | -0.1 | 0.15 | 1 | -0.34 |
| output | -0.23 | -0.28 | 0.43 | -0.14 | -0.085 | -0.028 | 0.14 | 0.42 | -0.44 | -0.43 | 0.35 | -0.39 | -0.34 | 1 |

Fig. 3. Correlation matrix.

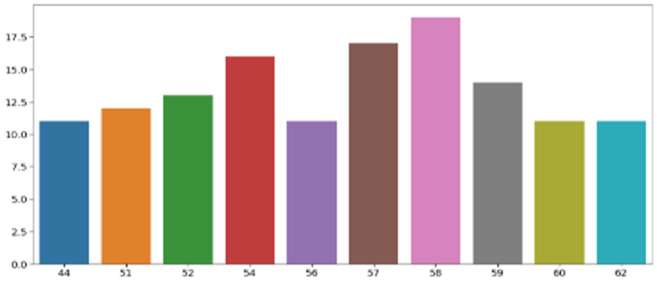


Fig. 4. Bar graph of age counts.

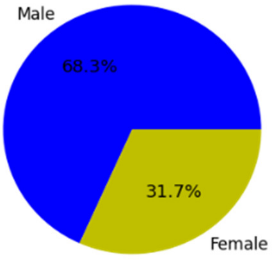


Fig. 5. Pie plotting for % of sex counts.

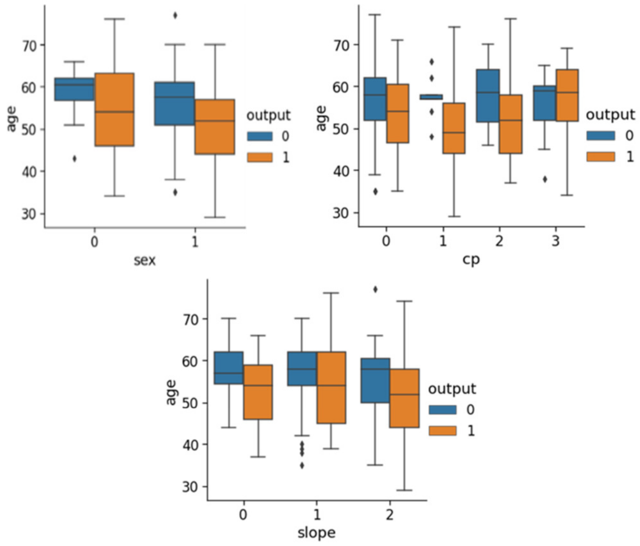


Fig. 6. Box plots of various attributes.

Figure 9 depicts the labeling count to the output (target) column in the entire data where 0 (no) heart attack has 140 records, and 1 (yes) heart attack has 163 records.

### B. Feature Selection Results

PCA was used to minimize data dimensions to build a model fast with less effort and increase efficiency [35]. Figure 10 depicts the extracted features after applying PCA for the 14 different features.

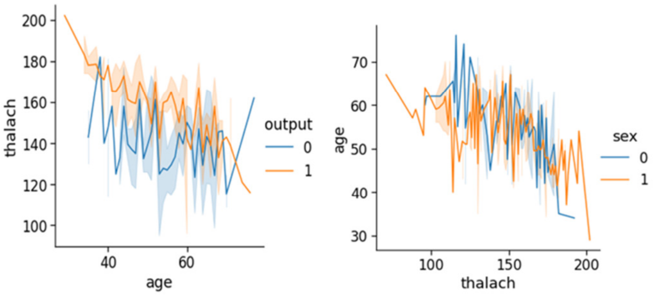


Fig. 7. Scatter plot between age, thalach, sex, and output.

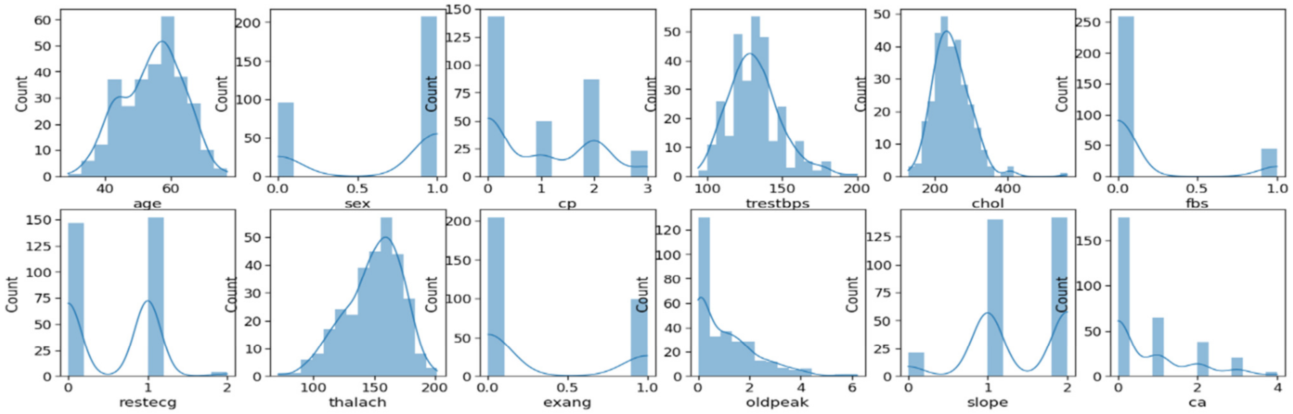


Fig. 8. Histogram of the heart disease dataset [24].

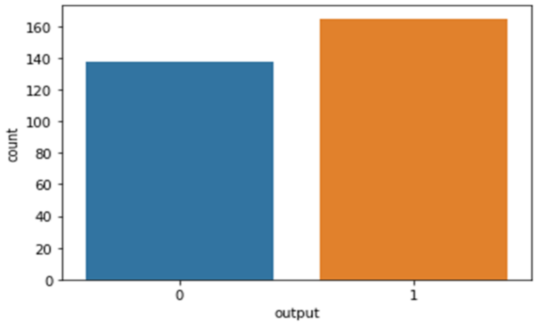


Fig. 9. Data labeling count of heart disease outcomes.

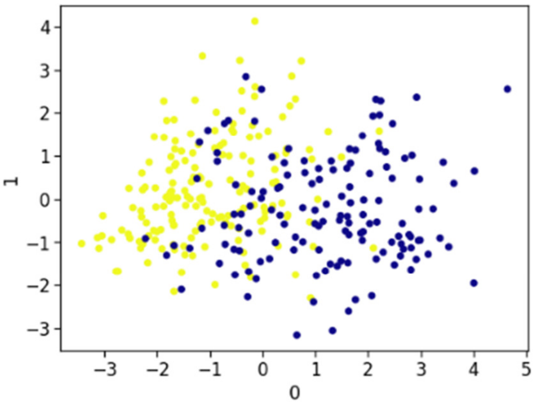


Fig. 10. Feature selection using PCA.

### C. Confusion Matrices

Figure 11 shows the confusion matrices for all four ML classifiers. These are binary confusion matrices for telling us whether a person suffers from a heart disease (1) or not (0). Figure 11 depicts the confusion matrix plots for the four classifiers after applying them with different hyperparameter settings. Figure 11(a) depicts the testing confusion matrix for the AdaBoost classifier in which TP instances were 33, misclassified instances were 3 FN and 2 FP, and TN instances were 23. Figure 11(b) depicts the testing confusion matrix for GB with 32 TP, 4 FN and 3 FP, and 22 TN. Figure 11(c) depicts the testing confusion matrix for ET, with 30 TP, 6 FN and 2 FP, and 23 TN. Figure 11(d) depicts the testing confusion matrix for CatBoost, with 32 TP instances, 4 for FN and 5 FP, and 20 TN [28, 36].

### D. Classification Results

Figure 12 depicts the classification performance metrics for (a) AdaBoost, (b) GB, (c) ET, and (d) CatBoost.

Figure 13 displays the precision-recall curve analysis for different models. The precision-recall curve is plotted between precision and recall, whereas the ROC curve analysis is plotted between TPR and FPR [37-39]. Figure 14 depicts comparative ROC curve plots for the classifiers. AdaBoost, GB, and ET achieved more than 0.9 AUC, but CatBoost achieved less than 0.9. The AdaBoost model was the most accurate predictive classifier for identifying heart attacks.

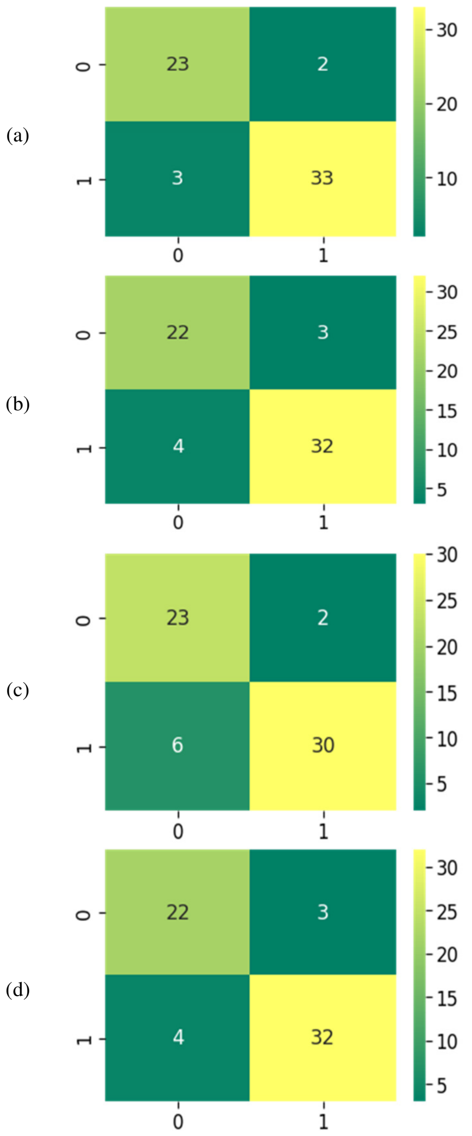
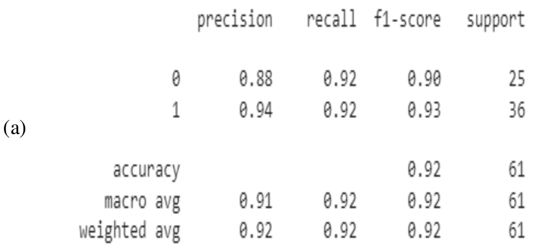


Fig. 11. Testing confusion matrices for (a) AdaBoost, (b) GB, (c) ET, and (d) CatBoost.

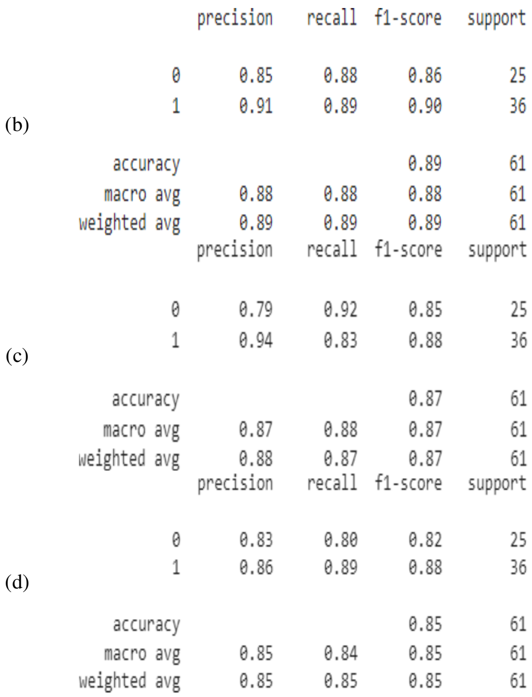


(a)

```
              precision    recall  f1-score   support

           0       0.88      0.92      0.90        25
           1       0.94      0.92      0.93        36

    accuracy                           0.92        61
   macro avg       0.91      0.92      0.92        61
weighted avg       0.92      0.92      0.92        61
```



(b)

```
              precision    recall  f1-score   support

           0       0.85      0.88      0.86        25
           1       0.91      0.89      0.90        36

    accuracy                           0.89        61
   macro avg       0.88      0.88      0.88        61
weighted avg       0.89      0.89      0.89        61
```

(c)

```
              precision    recall  f1-score   support

           0       0.79      0.92      0.85        25
           1       0.94      0.83      0.88        36

    accuracy                           0.87        61
   macro avg       0.87      0.88      0.87        61
weighted avg       0.88      0.87      0.87        61
```

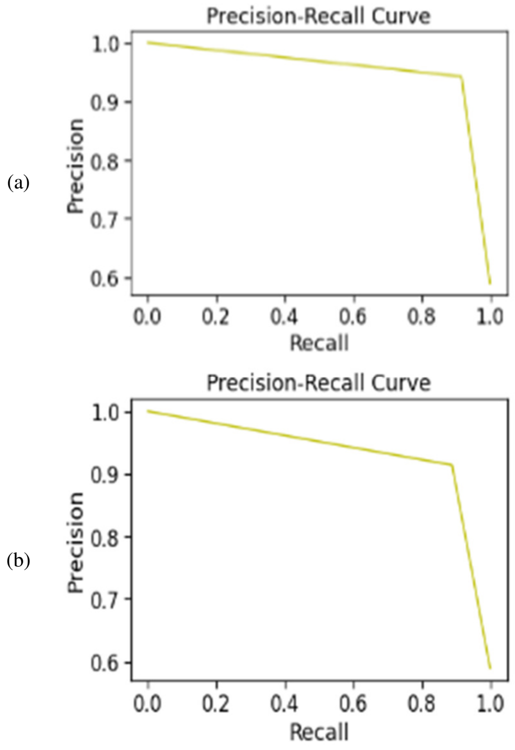
(d)

```
              precision    recall  f1-score   support

           0       0.83      0.80      0.82        25
           1       0.86      0.89      0.88        36

    accuracy                           0.85        61
   macro avg       0.85      0.84      0.85        61
weighted avg       0.85      0.85      0.85        61
```

Fig. 12. Performance metrics for (a) AdaBoost, (b) GB, (c) ET, and (d) CatBoost.



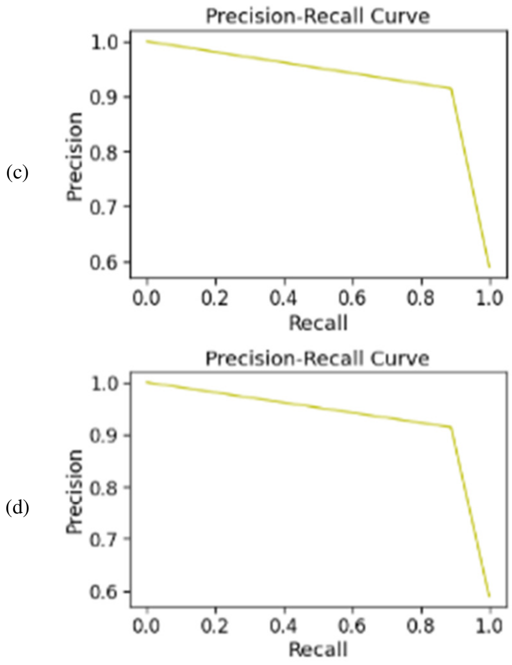


Fig. 13. Precision-recall curve plots for (a) AdaBoost, (b) GB, (c) ET, and (d) CatBoost.

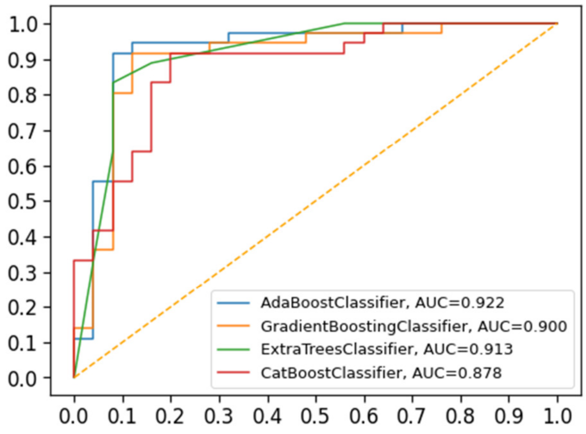


Fig. 14. ROC curve analysis for multiple classifiers.

## VI. DISCUSSION

Given the high frequency and financial cost of heart disease, the focus of the study is aligned with current healthcare goals. The performance of four different classifiers was compared, with particular attention paid to confusion matrices, ROC-AUC diagrams, and precision-recall curve diagrams. The AdaBoost ML classifier outperforms the competition with remarkable accuracy, F1 score, and precision measures. The results show that AdaBoost can accurately diagnose CHD, while practicing in real time has huge therapeutic intervention benefits. AdaBoost achieved the highest AUC value and accuracy among the models compared. However, further studies are required to verify the feasibility and effectiveness of the proposed model to serve other populations of patients as well. This study focused on generating viable CHD predictions, focusing on data preprocessing and feature selection. The proposed approach is highly effective in CHD diagnostics during the prediagnosis stage, and is characterized by better performance indicators for speed, accuracy, and feature selection for practical use.

## VII. CONCLUSION

Disease prediction is important because it enables physicians to become familiar with the dangers associated with certain diseases in their entirety. ML methods, including AdaBoost, have a lot of potential to improve classification accuracy. Given the increase in cardiovascular diseases and their long-lasting and damaging effects on the population, members of the healthcare industry look for ways to identify these diseases early and manage their manifestations. Classification performance was measured using ROC-AUC and precision-recall curves, where AdaBoost achieved the highest average accuracy, F1 score, and AUC. However, these findings require further investigation to generalize to other healthcare settings, testing the effectiveness of the proposed models for real-time identification of cardiac complications. However, some limitations have been observed, such as data fusion, bias issues, and the requirement for large clinical trials. Future studies should address these limitations, resolve ethical issues, and improve the applicability and reliability of the proposed model.

The potential of an ML-enabled device to predict CHD is apparent. Utilizing more contemporary prescriptive approaches, such as DL, can increase the likelihood of using chordate data along with previous medical records for greater efficiency. There are certain therapeutic interventions and surveillance for each patient's risk profile that could reduce healthcare costs. Furthermore, after considering ethical issues, patients could be involved in this process through simple applications.

## REFERENCES

[1] C. Bemando, E. Miranda, and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, Pekan, Malaysia, Aug. 2021, pp. 232–237, https://doi.org/10.1109/ICSECS52883.2021.00049.

[2] Y. Khourdifi and M. Bahaj, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 242–252, Feb. 2019, https://doi.org/10.22266/ijies2019.0228.24.

[3] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *Journal of Biomedical Informatics*, vol. 58, pp. S171–S182, Dec. 2015, https://doi.org/10.1016/j.jbi.2015.09.006.

[4] H. Kim, M. I. M. Ishag, M. Piao, T. Kwon, and K. H. Ryu, "A Data Mining Approach for Cardiovascular Disease Diagnosis Using Heart Rate Variability and Images of Carotid Arteries," *Symmetry*, vol. 8, no. 6, Jun. 2016, Art. no. 47, https://doi.org/10.3390/sym8060047.

[5] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, Jan. 2020, Art. no. 100330, https://doi.org/10.1016/j.imu.2020.100330.

[6] M. J. Jassim Ghrabat, G. Ma, and C. Cheng, "Towards Efficient for Learning Model Image Retrieval," in *2018 14th International*

*Conference on Semantics, Knowledge and Grids (SKG)*, Guangzhou, China, Sep. 2018, pp. 92–99, https://doi.org/10.1109/SKG.2018.00020.

[7] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, https://doi.org/10.1007/s12525-021-00475-2.

[8] M. Rolínek, D. Zietlow, and G. Martius, "Variational Autoencoders Pursue PCA Directions (by Accident)," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 12398–12407, https://doi.org/10.1109/CVPR.2019.01269.

[9] K. Dissanayake and M. G. Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, no. 1, 2021, Art. no. 5581806, https://doi.org/10.1155/2021/5581806.

[10] J. Abdollahi and B. Nouri-Moghaddam, "A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation," *Iran Journal of Computer Science*, vol. 5, no. 3, pp. 229–246, Sep. 2022, https://doi.org/10.1007/s42044-022-00104-x.

[11] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," *BMC Bioinformatics*, vol. 21, no. 1, Jul. 2020, Art. no. 278, https://doi.org/10.1186/s12859-020-03626-y.

[12] T. K. Sajja and H. K. Kalluri, "A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network," *Revue d'Intelligence Artificielle*, vol. 34, no. 5, pp. 601–606, Nov. 2020, https://doi.org/10.18280/ria.340510.

[13] S. N. Pasha, D. Ramesh, S. Mohmmad, A. Harshavardhan, and Shabana, "Cardiovascular disease prediction using deep learning techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 981, no. 2, Sep. 2020, Art. no. 022006, https://doi.org/10.1088/1757-899X/981/2/022006.

[14] C. Xiao, Y. Li, and Y. Jiang, "Heart Coronary Artery Segmentation and Disease Risk Warning Based on a Deep Learning Algorithm," *IEEE Access*, vol. 8, pp. 140108–140121, 2020, https://doi.org/10.1109/ACCESS.2020.3010800.

[15] K. Vayadande *et al.*, "Heart Disease Prediction using Machine Learning and Deep Learning Algorithms," in *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, Greater Noida, India, May 2022, pp. 393–401, https://doi.org/10.1109/CISES54857.2022.9844406.

[16] T. Amarbayasgalan, V. H. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets," *IEEE Access*, vol. 9, pp. 135210–135223, 2021, https://doi.org/10.1109/ACCESS.2021.3116974.

[17] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, 2021, Art. no. 8387680, https://doi.org/10.1155/2021/8387680.

[18] S. Prusty, S. Patnaik, and S. Kumar Dash, "Comparative analysis and prediction of coronary heart disease," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 2, Aug. 2022, Art. no. 944, https://doi.org/10.11591/ijeecs.v27.i2.pp944-953.

[19] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked*, vol. 26, Jan. 2021, Art. no. 100655, https://doi.org/10.1016/j.imu.2021.100655.

[20] M. Sebastiani, C. Vacchi, A. Manfredi, and G. Cassone, "Personalized Medicine and Machine Learning: A Roadmap for the Future," *Journal of Clinical Medicine*, vol. 11, no. 14, Jan. 2022, Art. no. 4110, https://doi.org/10.3390/jcm11144110.

[21] F. Özbilgin, Ç. Kurnaz, and E. Aydın, "Prediction of Coronary Artery Disease Using Machine Learning Techniques with Iris Analysis," *Diagnostics*, vol. 13, no. 6, Jan. 2023, Art. no. 1081, https://doi.org/10.3390/diagnostics13061081.

[22] C. Eyupoglu and O. Karakuş, "Novel CAD Diagnosis Method Based on Search, PCA, and AdaBoostM1 Techniques," *Journal of Clinical Medicine*, vol. 13, no. 10, Jan. 2024, Art. no. 2868, https://doi.org/10.3390/jcm13102868.

[23] G. Deivendran, S. Vishal Balaji, B. Paramasivan, and S. Vimal, "Coronary Illness Prediction Using the AdaBoost Algorithm," in *Sensor Data Analysis and Management*, John Wiley & Sons, Ltd, 2021, pp. 161–172.

[24] D. Lapp, "Heart Disease Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[25] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease." UCI Machine Learning Repository, 1989, https://doi.org/10.24432/C52P4X.

[26] A. Dey, J. Singh, and N. Singh, "Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis," *International Journal of Computer Applications*, vol. 140, no. 2, pp. 27–31, Apr. 2016, https://doi.org/10.5120/ijca2016909231.

[27] J. Tang, A. Henderson, and P. Gardner, "Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets," *Analyst*, vol. 146, no. 19, pp. 5880–5891, 2021, https://doi.org/10.1039/D0AN02155E.

[28] P. Theerthagiri, "Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique," *Intelligent Systems with Applications*, vol. 16, Nov. 2022, Art. no. 200121, https://doi.org/10.1016/j.iswa.2022.200121.

[29] R. Shafique, A. Mehmood, S. Ullah, and G. S. Choi, "Cardiovascular Disease Prediction System Using Extra Trees Classifier." Research Square, Sep. 16, 2019, https://doi.org/10.21203/rs.2.14454/v1.

[30] X. Zhang *et al.*, "An accurate diagnosis of coronary heart disease by Catboost, with easily accessible data," *Journal of Physics: Conference Series*, vol. 1955, no. 1, Mar. 2021, Art. no. 012027, https://doi.org/10.1088/1742-6596/1955/1/012027.

[31] M. J. J. Ghrabat *et al.*, "Fully automated model on breast cancer classification using deep learning classifiers," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, Oct. 2022, Art. no. 183, https://doi.org/10.11591/ijeecs.v28.i1.pp183-191.

[32] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, https://doi.org/10.1109/ACCESS.2020.3001149.

[33] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, "Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms," in *2012 IEEE 12th International Conference on Data Mining Workshops*, Brussels, Belgium, Dec. 2012, pp. 9–16, https://doi.org/10.1109/ICDMW.2012.29.

[34] M. J. Kraak, "Exploratory Visualization," in *Encyclopedia of GIS*, S. Shekhar and H. Xiong, Eds. Boston, MA, USA: Springer US, 2008, pp. 301–307.

[35] S. Usha and S. Kanchana, "Predicting Heart Disease Using Feature Selection Techniques Based On Data Driven Approach," *Webology*, vol. 18, no. 4, pp. 97–108, 2021.

[36] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Scientific Reports*, vol. 10, no. 1, Nov. 2020, Art. no. 19747, https://doi.org/10.1038/s41598-020-76635-9.

[37] M. J. J. Ghrabat, G. Ma, Z. A. Abduljabbar, M. A. Al Sibahee, and S. J. Jassim, "Greedy Learning of Deep Boltzmann Machine (GDBM)'s Variance and Search Algorithm for Efficient Image Retrieval," *IEEE Access*, vol. 7, pp. 169142–169159, 2019, https://doi.org/10.1109/ACCESS.2019.2948266.

[38] M. J. Ghrabat, G. Ma, P. L. P. Avila, M. J. Jassim, and S. J. Jassim, "Content-based image retrieval of color, shape and texture by using novel multi-SVM classifier," *International Journal of Machine Learning and Computing*, vol. 9, no. 4, pp. 483–489, 2019.

[39] S. M. Alanazi and G. S. M. Khamis, "Optimizing Machine Learning Classifiers for Enhanced Cardiovascular Disease Prediction," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12911–12917, Feb. 2024, https://doi.org/10.48084/etasr.6684.