Utilising 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 College, Baghdad, 10013, Iraq
mudhafar.jalil@iips.icci.edu.iq

Siamand Hassan Mohialdin
College of Health Science, 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, 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
Mechatronics Department, Faculty of Engineering, Debrecen, University of Debrecen, 4028, Ótemető u. 4-5, Hungary
husam@eng.unideb.hu (corresponding author)

Received: 1 January 202X | Revised: 2 February 202X | Accepted: 3 March 202X ("ETASR dates" style)

 DOI: https://doi.org/10.48084/etasr.XXXX

ABSTRACT

Even though occurrence of heart disease (HD) is unavoidable, predictor is one of the phases that help prepare doctors to best be ready to develop a solution for their patients. Some of the likely challenges that health care professionals could encounter include; the fact that heart failure is a very prevalent condition in society, and the presentation of its symptoms may mimic the effects of aging. The use of machine learning (ML) for this problem encounters a significant obstacle: the case of having high dimension data where the extracted information is numerous. This paper aims at assessing a negative implication of predicting HD as this forms the paramount importance for enabling the healthcare doctors in making proper health care decisions that would enable the improvements of the patient survival. This has posed a major challenge in the forecast of HD hence; in this work we present a fresh perspective that is a combination of PCA and feature selection methods to help improve the prediction capacity of the method especially in the loadings aspect. In a case of the models of this study topic, it has been realized that feature selection was one of the components that contributed so much to the encouragement of success. Consequent to such discoveries, this research has put forward a method of data reduction to lessen the dimensionality of the findings then employed the features selection method to pinpoint significant attributes that tend to be attributable to heart diseases. Further, there cannot be an extraneous variable, or a process that happened in preparing the data, which tends to bias the result in some specific direction disturbing the source material. The data for this work has been downloaded from Kaggle and can also be found on the UCI ML Repository under the category heart disease. The kind of classifiers employed in the study were four while the characteristics of the features which included were thirteen and the label attribute of the study was one in kind. Specifically, when the PCA, Random Forest, and Decision Trees are combined with the AdaBoost classifier on the Cleveland dataset, an accuracy of 96.8% is achieved. Their selection was based on their proficiency in performing classification tasks, as well as their proven track record of performance in prior studies.

Have to experimental results demonstrated the efficacy of PCA, Random Forest and Decision Trees as it consistently produced superior results across most classifiers. Random Forest, Decision Trees and PCA's contribution is its capacity to substantially reduce dimensionality while preserving essential information. To validate in this proposed prediction model, the compared it to an existing method employing two classifiers: random forest and logistic regression. In this proposed model using the AdaBoost classifier outperformed other machine learning classifiers regarding precision, recall, and AUC value. In this study presents a robust method for predicting cardiac disease by effectively reducing dimensionality via Random Forest, Decision Trees and PCA and identifying crucial attributes via feature selection.

Keywords- random forest; decision trees; principal component analysis (PCA); machine learning (ML) classifiers; coronary heart disease (CHD); hyperparameters; prediction

I. INTRODUCTION

Cardiovascular disease continues to be the primary cause of mortality globally, responsible for around 18 million deaths each year. Although there have been notable advancements in medical technology, accurately forecasting cardiac disease continues to be a challenging task. Existing prediction approaches often demonstrate subpar performance, resulting in missed diagnosis and postponed actions. Coronary heart disease (CHD), often known as cardiovascular disease (CVD), is the leading cause of death worldwide. As a result, several researches have been conducted to predict the early detection of cardiac issues and identify the most important risk factors connected with the condition.

Despite extensive attempts, the forecast's accuracy has remained inadequate, and pinpointing the most influential risk factors has been challenging [1]. Extensive study has been conducted for many years on the early diagnosis of this particular kind of sickness. Data analytics approaches have been used to assist healthcare workers in detecting early indications of cardiac disease.

Prospective patients may undergo many diagnostics to mitigate the burden of experiencing such a condition [2-3]. Accurate methods for predicting the early stages of heart disease, as proposed in this work, may be crucial for preserving people's lives. Numerous types of research have already been undertaken to forecast the risk of CHD. Deep learning (DL) is transforming various applications in various industries. Through image analysis and predictive modelling, deep learning improves medical evaluation and treatment planning in healthcare. The automotive industry is utilizing DL for self-driving systems, which will make transport safer and more efficient. DL is enhancing fraud detection and algorithmic trading tactics in finance. Retailers are using 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. These developments demonstrate the broad influence of deep learning in industries such as healthcare, automotive, finance, retail, and technology & ML (Machine Learning) techniques. ML-based methods mostly suggested single or ensemble classification techniques, using feature extraction/selection methods to enhance results [4-6], Deep learning algorithms are now being applied extensively to detect CHD. Most available approaches divide an experimentation data set into two portions for training and testing. Next, they employ classification techniques to create prediction models from entire or randomly picked training samples. Consequently, models are much more likely to be fitted on consistently distributed datasets & mislabel unevenly distributed (biased) data [7].

Common machine learning classifiers like Adaboost, Gradient Boosting, Extra Trees, and CatBoost are employed with tried-and-true methods like PCA for feature reduction. However, it does bring up some interesting new points.

Deep Learning (DL) has received worldwide appreciation for its disruptive influence across multiple domains. The quantity of applications in deep learning has lately surged to a level where it has shown its own capability and adaptability, exemplified by models such as DCNNBT - an early classifier for brain tumours based on deep convolutional neural networks. Using deep neural networks, this model has

generated a more precise and detailed image, received on August 31st, 2017, compared to what can be seen by humans using approaches that just depend on subtle patterns detected by convolutional filters. In addition to that, the use of data augmentation along with transfer learning made the model more flexible and the detection of the brain tumour was more accurate. These approaches are also fairly easy to scale and have good fault tolerance and as such are well suited for large scale use. A rather practical improvement based on U-Net-style models in image segmentation is a leap forward towards better definition of the pathological and structural margins.

It has also helped in improving the capacity of the clinician to arrive at a sound judgement when it comes to treatment and diagnosis. It is, thus, not just for health as the consequences extend to other areas and are not limited to the establishment of the most recent methodologies for processing the outcomes of biological experiments. For example, it is possible to use DL in cancer diagnosis, for example, the method based on ideas for predicting the efficacy of anti-angiogenic drugs described in this paper. The milestones illustrate the extensive scope and profound influence of DL capabilities in tackling wide-ranging, intricate issues in several fields, highlighting its disruptive potential for innovation in solving practical concerns [8].

According to Wikipedia, PCA (principle Component Analysis) 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 principle components, after identifying the principle components, which are the directions in the data that have the most variation.

The research was done in the following manner. Pre-diagnosis of coronary artery disease (CAD), the system simply records five essential factors: Age, hypertension, usual chest pain, T wave inversion, and localised wall motion anomalies. This approach uses a blend of eight search techniques to find relevant features. It then employs Principal Component Analysis (PCA) and an AdaBoost algorithm for the classification task. This is the most exceptional performance recorded on a publicly available dataset with such a limited number of features. This strategy enhanced efficiency, precision, and resilience compared to prior methods that relied on a wider array of qualities. This feature makes it a viable instrument for the early identification of CAD, since it specifically addresses certain obstacles such as data integration, algorithmic bias, and clinical validation [42].
Distinctive attributes for CAD diagnosis differentiation compared to the previous method.

In contrast to prior techniques, our method distinguishes itself by using a multitude of characteristics and various types of features to identify coronary artery disorders (CAD). What is the precise definition of key drivers?

Reduced Feature Count: In contrast to previous research that used a range of 16 to 40 features, this technique just requires five feature components. The factors that contribute to this condition are ageing, hypertension, usual chest discomfort, T-wave inversion, and regional wall motion abnormalities. By lowering the amount of features, your model becomes less complicated and more efficient, while yet keeping accuracy. The proposed method incorporates eight feature selection-based search techniques, including evolutionary, best first genetic harmony particle swarm optimisation, greedy stepwise_RANKmongeneous evolution. This comprehensive trading system ensures that the selected features align with the desired criteria, thus preprocessing effectively prepares the data for the subsequent stage. PCA, Principal Component Analysis, is a technique used to reduce the dimensionality of data by transforming it into a new space where the most important aspects are emphasised. This not only decreases the computational cost of the dataset, but also improves correspondingly as the amount of useless data decreases.

The suggested methodology utilises the PCA and DT algorithms for classification, which is a technique in ensemble learning that combines numerous weak classifiers to create a powerful classifier. This method significantly enhances the performance of the model in comparison to all conventional single classifiers documented in prior work.

Outcome: The approach achieves a classification accuracy of 96.8% on the https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset. Data Testing: accuracy achieved highest proposed. The White wine dataset has the shortest feature vector sizes. Prior research has shown that increasing the number of features improves scalability, but it has been observed that the accuracy levels achieved are either comparable or worse. Therefore, we conducted tests in the past to investigate this further. The present research is based on the findings of those previous investigations.

It provides a superior and more pragmatic approach for early detection of CAD compared to the current approaches. The process involves using unsupervised feature selection and classification algorithms to obtain a limited number of highly significant characteristics.

Regarding this great challenge, the improved method to identify CHD has been proposed in this research that potentiates development of chances of correct prediction with the help of utilization of PCA and feature selection. The all novel procedure provided a feasible modality to the practical management of the circadian timing disturbance.

The method used is Principle Component Analysis (PCA) which will enhance the number of characteristic in the prediction of the HD in the given dataset of 13 to superimpose principle components. Also, it is necessary to consider the fact that often some of the attributes are not unloaded. This helps to minimize the computational costs of using machine learning classifiers and at the same time, enhances the classification's likelihood of finding a way of generalizing his search results. In feature selection it is the process of selecting a small presumably relevant set of features from the complete set of features. In machine learning, it is least of preference to have irrelevant or features that are beyond necessary since they are always a hindrance.

Context: Applying these techniques in the prediction of CHD means that health specialists have been trained on how

to handle the special issues associated with Cardiovascular and Health Data.

However, to enhance the overall efficiency of the above-discussed prediction model in CHD easier, the following directions require attention: tuning up the parameters used for the PCA analysis and looking at other methods of the dimension reduction. Such changes enable one to get a model that gives the best results in a specific domain of application.

Finally, this paper evaluated the efficiency of the four types of machine learning classifiers for CHD diagnostics and came to a conclusion on the choice of the classifier that should be used. Moreover, it is worthy to note that the mission-oriented attitude of this operation is considered to be highly effective to search the increase in the efficiency in cooperation with CHD.

Despite the researcher maybe have overlooked them in this proposal, the following have provided novel findings and trends which are relevant for identifying CHD. This may include features which were important ranked highly, models which can be interpreted and new relations which weren't considered in the solution.

Ultimately, the research highlights a significant area where precise prediction of CHD may be used. When coupled with prompt human interventions and customised treatment procedures, this approach has the potential to not only save lives but also save healthcare costs. This practical application showcases a significant and innovative result of the study.

This study proposes the integration of a dimensionality reduction strategy and several categorization models to achieve two primary goals: (1) effectively representing the characteristics of the dataset, and (2) using machine learning methods to help in producing accurate predictions. The research used the HD Prediction dataset from the Kaggle Repository. Various classification models, including Gradient Boosting, AdaBoost Classifier, Catboost classifier, and Extra tree classifier, were employed with specific hyperparameter values. The exploratory data is visualized using several plots, and the data is pre-processed using a standard scaler.

The features are chosen, and the PCA feature dimensionality reduction approach is used. Finally, a thorough comparison of various classifiers for classifying HD to accurately predict heart disease is performed.

This article's primary contributions are as follows:

1) To study several ML classifiers and classification techniques for HD detection & prediction.
2) To prepare the data using pre-processing techniques to solve the incompleteness and unreliability problem of the HD dataset.
3) Use the PCA feature selection method for extracting the most valuable features to enhance classification results by dimensionality reduction.
4) It provides increased classification and prediction results using a prediction model by implementing different machine learning classifiers.
5) Comparing advanced machine learning classifiers with existing classification methods according to performance measures, among others, & predicting whether a patient has HD.

A remainder of a paper has a following structure: Section II studies various existing ML classification techniques in disease prediction, especially heart disease. A new HD prediction model is offered in section III. Section IV and V compares and contrasts a proposed model's results and findings. At last, discussion and concluding remark for this work is provided with future work suggestions in section VI and VII.

II. LITERATURE REVIEW

This section explains research methods to identify and identify various cancers in different medicines. Numerous research works have been carried out to forecast coronary heart disease by employing different kinds of machine learning techniques using feature selection approaches [9-14] Still, recognition of many features is more difficult and large feature dimensions exceed the training time for the model. For this purpose, in the [1] study, a prediction model of CVD (Coronary Artery Disease) was proposed using supervised machine learning methods. These methods are referred to as Bernoulli Naive Bayes (BNB) algorithm, Random Forest (RF) algorithm, & Gaussian Naive Bayes (GNB) algorithm. In addition to that, several risk factors that are related to diseases are discussed in this article. It did this by utilizing the huge data set accessible in the Cleveland database of the ML repository at the University of California, Irvine. This database contained information regarding heart disease patients. Results demonstrated that the accuracy of the GNB and BNB models was 85%, while the accuracy of the RF model was 75%. Furthermore, the accuracy, recall, & f1-score of GNB & BNB were higher compared to RF, stating their significance in figuring out how to predict the early diagnosis of the situation. This was the case because the Random Forest used a randomization technique.

HD prediction has also been completed with DL techniques [15-16] for both feature selection and classifications for available HD datasets. In [17], a CNN model 3D U-net was proposed with some improvements for segmenting the coronary artery to forecast heart attack risk. This model was implemented with the assistance of various datasets, and it had two contexts: the first context did not have a centerline, and the second context did have a centerline. They can accomplish this by utilizing a novel local feature to acquire the data about the ventricles & utilize a Deep Belief Network (DBN) to collect features required to regress outline dimensions of the biventricular. Both of these methods are combined. A set of features, DBNs, and training regression networks enables it to retrieve high-level information. It makes it possible to correctly distinguish the left-side and right-side ventricles while only requiring moderate computing power. This is a significant advancement in the field of AI. Results of these types of trials indicated that the best possible impact could be obtained whenever the dice coefficient reached 0.8291 & total accuracy of the model was obtained at 78.782 percent, the highest among the three models. FCN was used for image

segmentation, and both the testing and training networks utilized the natural image dataset PASCAL VOC2012. However, FCN is not without its limitations; the obtained results are still quite fuzzy & smooth, and they are not particularly attuned to the visual details. The effectiveness of DL in image segmentation has significantly impressed individuals, although it is not appropriate for more complex medical images.

Also, both ML & DL techniques work together to predict CHD risks [18-19] developed an effective strategy for forecasting the risk of CHD using 2 DNNs trained on datasets arranged in a logically consistent manner. The well-run training sets make it possible to construct higher-precision prediction models. Korean National Health & Nutritional Examination Survey provided the dataset employed for this investigation. On the dataset, they performed two different kinds of tests. The first one demonstrated how well the suggested method's PCA and variational Autoencoder models might increase the overall performance of a single DNN. In experiment 2, a comparison was made between the suggested approach and other ML methods already used. The outcomes of the experiments indicated that the suggested technique is superior to traditional machine learning techniques, as it achieves higher levels of accuracy (0.892), specificity (0.840), precision (0.911), recall (0.920), f-measure (0.915), and area under the curve (0.882).

An accurate diagnosis of cardiac illness might minimize possible risks to one's life, whereas an inaccurate diagnosis can have the opposite effect and end in death. The findings and analyses of the UCI machine learning HD dataset are compared & contrasted in the article [20] by applying various ML techniques, including DL. To enhance the validity of findings, the dataset contains certain unimportant attributes, which are removed with an isolation forest. In addition, the data are normalized. An accuracy of 94.2% was achieved with the application of the DL technique. The issue that has arisen is that the sample size of the dataset is small. Thus, the results of DL and ML may improve substantially if a large amount of data is provided to the models.

Although doing further modifications on say a large data set, if we use the Deep Learning technique, we can predict still better results. This shall be achieved through the following: proposing a study therein it shall be posited that the area of CHD prediction shall benefit from the use of machine learning methodologies thereby enhancing the reliability of its predictions. On models that we explored we have Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM). Finally an optimization of the parameters was done using grid search where three iterations and 10-fold cross-validation was used. Consequently, the parameters that were used in the evaluation of the model included Accuracy as well as F1 Score among other parameters. Sensitivity, specificity, positive predictive values, negative predictive values are four very large and important parameters, which help in deciding the results of the classification model. All these metrics may be arrived at the help of a confusion matrix. Such kind of metrics is applicable to cases where two variables are compared in order to establish some relationship between them. RF model led to consider that the classification accuracy corresponds to 0.

The proposed model is applauded more than the SVM and LR models with the accuracy of 0. 929. Thus, at the end of this study, the authors purposely selected the RF model, which is the centroid of the highest sensitivity, to acknowledge that there is no heart illness. To sum up for this research, it can be concluded that the RF model of this research is sensitive and specific in evading CHD in Yi people. Specificity was estimated to be 54% while the sensitivity was measured to be 98% the PPV of the method however was found to be 29 while the NPV was also 29.15% and 98.68% respectively. This has a positive implication on its use in diagnosing Clinic CHD early because there is no good measurement technique up to date [38].

Therefore, utilizing the machine learning for approach, Mahajan et al were able to achieve a definite detection of the Coronary heart diseases (CHD) in this case. It was observed that out of the developed three models namely, RF, LR and SVM, the effectiveness of all the models was compared. Therefore, we compiled a comprehensive dataset on rescue heart disease by using a diverse range of information sources. There are 281 female and 909 male instances in the real-world dataset, together with patient age, sex chest pain kind, resting blood pressure (mm Hg), serum Cholestroral columun (parents' cholesterol level), etc. All models were optimised and tested by repeated cross-validation to prevent overfitting. Performance measures included Accuracy, Specificity (Sp), Sensitivity (Se), F1-Score, Positive Predictive Value (PPV), and Negative Predictive Value. With the 0.929 accuracy, RF algorithm was the best, followed by SVM and LR [39].

This study will use machine learning to identify coronary artery disease (CAD). In this experiment, we select five features (age, hypertension, typical chest pain, T-wave inversion, and regional wall motion abnormality) and integrate eight search algorithms: Naive Bayes, Multilayer Perceptron (MLP), Decision Table, JRip, Ridor, X-2, HoeffdingTree, Bayesian Network, and PCA methods with the AdaBoostM1 semiclassical learning paradigm. On Z-Alizadeh Sani dataset, our technique achieved 91.8% classification accuracy, the highest recorded result with the fewest characteristics, they stated. Early and reliable CAD identification might help doctors respond faster, improving patient care and survival [40].

The research introduces iris image-based noninvasive coronary artery disease (CAD) diagnosis. Using iridology and advanced image processing, the research produced iris texture patterns from 198 volunteers—94 with CAD and 104 without. Wavelet transform, GLCM, and GLRLM extract features from pre-processed pictures to segment the iris heart. Using several machine learning classifiers, the SVM achieved 93% accuracy. Telemedicine applications like telediagnosis may benefit from this method [41].

But there are also a lot of problems. The segmentation and lesion detection process begins with the identification and positioning of coronary arteries. Manually identifying

coronary arteries is time-consuming, and the operator's biases might impact the segmentation findings. Second, medical pictures have certain qualities. As a first point, it is hard to tell whether a medical picture 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 were observed, such as:

1) Predictive models do not learn effectively from real-world heart disease datasets since each contains an imbalanced percentage with varying variations from the rest of the data.
2) It was also discovered that the accuracy gained is insufficient when a model is tested for real-world data problems, which may differ significantly from the dataset on which it was trained. That dataset must be normalized to prevent the training model from being overfitted.
3) Most currently accessible prediction models were trained to utilize the entire training set or a randomly
4) Selected subset. Several approaches have been implemented on the popular dataset Cleveland, although the accuracy achieved by all of them is highly dependent on time calculations.
5) A strategy for generating training data by discriminating between ordinary and extremely biased subgroups has not been developed to create a reliable prediction model.
6) The difficulty was the sample size of the dataset was not huge, which caused the results for deep learning not to do so well.

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

Problem	Existing Approach	Proposed Approach
Imbalanced datasets	Traditional models struggle to learn from imbalanced datasets	The proposed method utilises a hybrid strategy that integrates oversampling, undersampling, and synthetic minority oversampling strategies to tackle the uneven character of the dataset.
Overfitting and generalizability	Models trained on entire training sets or random subsets may overfit and lack generalizability to real-world data	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 technique employs a fusion of principal component analysis (PCA) and correlation analysis to detect the most significant and useful characteristics.
Sample size limitations	Small sample sizes can negatively impact the performance of deep learning algorithms	Proposed method employs a transfer learning approach to leverage pre-trained deep learning models and improve performance with limited data.

In Table I summarize the key differences between the proposed approach and existing approaches to coronary heart disease (CHD) detection. It highlights the challenges faced by existing approaches and how the proposed method addresses these challenges.

The literature review revealed that basic data augmentation could be adversative to the test dataset. The coronary heart segment's performance will improve more if the training dataset's quality is raised. As a result, several types of feature identification have become harder. There will be a lack of features and hardly noticeable variations in practice. Third, there are several potential sources of disruption during the collection of medical data, including but not limited to variations in patients, equipment, variables, as well as running environments.

Based on the above reasons, the have presented method resolved these issues. Therefore, this article focuses on the CHD dataset, Cleveland, studies feature selection methods in the medical field with ML, & optimizes algorithms following clinical application needs. The dataset used in this paper and the related data processing are presented. PCA was used to reduce feature dimensions to get the dataset's best features for coronary arteries. The results of this experiment are contrasted and examined in light of the assessment criteria for the machine learning outcomes and the outcomes of other methods. The ease of diagnosing and treating CADs may provide clinicians with the greatest predictive model available.

III. PROPOSED APPROACH

In this section, a heart disease forecasting model is developed that follows some procedures by applying different ML classifiers to solve an identified research gap in early coronary prediction with the proposed solution discussed below.

A. Statement of the problem

There is a research gap in early coronary prediction, whether HD or not. Early prediction of coronary artery disease using clinical criteria is lacking in studies. Daily plaque buildup lowers cardiac blood flow. This refers to the early stage of coronary artery disease. Plaque is composed of cholesterol walls and chemicals.

Heart oxygen levels are progressively falling. The valve and wall will shrink, resulting in a different look from the reference picture. Present a method to detect early coronary artery disease with greater accuracy [21]. Heart disease is a dangerous disease on the rise in both developed and developing countries.

Early and precise diagnosis of this condition is crucial for averting additional harm to patients and preservatives their lives [22]. Although it has drawbacks among standard invasive-based procedures, angiography is the most well-known method for identifying heart issues. In contrast, non-invasive technologies, such as artificial learning-based computational algorithms, are deemed more reliable and effective for identifying cardiac disease [13]. This article offers an intelligent computational prediction model for heart disease

detection & diagnosis by testing different ML classifiers in this research.

B. Proposed methodology

One of the primary causes of HD is a variety of conditions affecting the heart. Blood vessel diseases, like coronary disease, are included in HD. Arrhythmias (irregular heartbeats), etc. This study's primary goal is to forecast CHD. This work gathered HD data from Kaggle and then performed pre-processing to clean the data, including some pre-processing techniques like missing value check, standard scaling, etc. Once data pre-processing is done, essential features are selected using the PCA technique is used as a feature selection technique to reduce the data dimensions that are helpful to speed up the training of the model. Then, data is separated into training & testing sets to conduct experiments. Finally, classification is done by applying four (gradient boosting, AdaBoost classifier, CatBoost classifier, & extra tree classifier with their hyperparameter settings) machine learning classifiers and predicting heart disease. These classifiers evaluated performance using F1-score, Accuracy, Precision, and Recall parameters. Also, finally, prediction is provided using confusion matrices for the prediction model. Figure 1 depicts a block diagram of this proposed prediction model to predict heart disease.

The prediction model follows the following steps to implement it. These are discussed below.

1) Dataset description

Data gathering or collection is the first step of any prediction model to experiment on the collected dataset. The dataset, "Heart Disease Dataset," was obtained via Kaggle. The dataset is accessible at the following link: "https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset." For this work, data are retrieved from the Kaggle repository database. The Heart dataset comprises 14 attributes, most integer & float datatypes. The dataset comprises four databases with 76 attributes: Long Beach V, Cleveland, Hungary, and Switzerland. However, the research concentrates primarily on fourteen key characteristics. The primary target attribute indicates the presence of heart disease (0 for absence of disease, 1 for disease).

Age, gender, 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 insight into patients' health and heart condition. The dataset is useful for researching the correlation between these characteristics and cardiovascular disease while protecting the patient's privacy.

2) Data preparation/preprocessing

Preprocessing is a technique used to get data that is comprehensive, coherent, and easily understandable. The mining results generated using machine learning algorithms are influenced by the quality of the data. High-quality data leads to informed and accurate decision-making. Consequently, the FHS dataset is included by means of the subsequent preprocessing procedures.

• Insignificant attributes might impair the model's performance and lower the learning rate. Feature selection is a crucial part of preprocessing, since it involves selecting the characteristics that have the most impact on predicting the intended outcomes. Utilizing an automated feature selection in the dataset would have resulted in the elimination of crucial characteristics. Hence, using an analytical method yields superior results. The most important step is to prepare a dataset by following several mechanisms: information is checked, null values are checked, then describing the data summary is, and a standard scaler is used to standardize the input features. If there is any null value present, fill it with zero. Nonetheless, there is no null value present in the collected dataset. The standard score of each feature is measured by: $= (x - mean) / SD$ (1)

When the mean of the training feature is '0', the sample's standard deviation is '1', and the mean is False if the value of with_std is False. A StandardScaler instance is set with default hyperparameters.

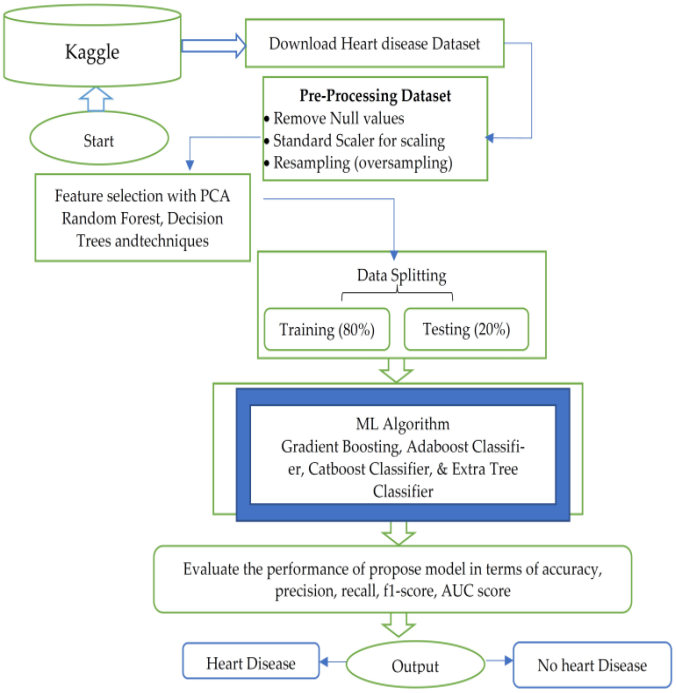


Fig. 1. Block Diagram of Proposed HD Prediction Model.

The advantages of using the method prposal

Principal Component Analysis (PCA) in conjunction with random forests and decision trees for early detection of Coronary Artery Disease (CAD) have been established in many studies, listing them as follows:

- Recorded data exhibits improved performance when Principal Component Analysis (PCA) is combined with Random Forests and Decision Trees. This approach is particularly useful for identifying the first CAD cases.
- PCA, or Principal Component Analysis, is a technique that decreases the number of dimensions in a dataset

and converts it into a new space while retaining crucial characteristics. It reduces the computational load and enhances the performance of Random Forest and Decision Tree models.

- Enhanced Model Performance: The performance of Random Forest and Decision Tree classifiers is enhanced by picking the top features that provide the best explanatory power for the nodes using PCA. This combination yields significantly improved accuracy, precision, recall, F-measure, AUC, and MCC compared to using classifiers without PCA.
- Manages Multicollinearity: By using Principal Component Analysis (PCA), it is possible to mitigate the issue of multicollinearity by generating uncorrelated principal components.
- As a potent tool for enhancing patient care. It is preferable to ensure that models such as Random Forest or Decision Tree are in a stable state, since they have a tendency to fail when dealing with correlated information.
- Enhanced Interpretability: This has the potential to improve the interpretability of a model by reducing the complexity of a dataset with several characteristics. Utilising components such as Decision Trees may provide a distinct set of decision criteria based on the major attributes, hence enhancing the probability that physicians would find it simpler to comprehend and have confidence in the predictions generated by our model.
- When using Random Forests and Decision Trees with PCA, Principal Component Analysis may effectively identify the most significant features, making it the optimal approach for feature selection. The consequence of only these components is diminished. Consequently, this leads to more accurate and exact predictions.
- The combination of Principal Component Analysis (PCA) with AdaBoost and Random Forest classifiers yielded the best level of accuracy in the classification task, particularly when using a limited amount of features. Efficiency is crucial since it enables quick and accurate diagnosis, facilitating timely treatment.
- Applying Principal Component Analysis (PCA) in combination with Random Forests and Decision Trees may effectively identify Coronary Artery Disease at an early stage, serving as a significant approach for preventative machine learning methods.

3) The selection feature of Principal Component Analysis (PCA), Decision Trees, and Random Forest and techniques.

Feature selection is an essential and critical stage in enhancing prediction models for cardiovascular disease. This research employs Principal Component Analysis (PCA), which is one of the primary approaches used for dimensionality reduction. Principal Component Analysis (PCA) enables the transformation of data into a new coordinate system, while retaining important information about the data. This, in turn, simplifies the analysis process. Principal Component Analysis (PCA), a commonly used method for feature selection, significantly reduces the dimensionality of the feature space to improve computational stability. It is particularly useful for understanding correlated information, which is crucial in predicting cardiac disease.

While PCA might serve as an alternative option It should be noted that there would be some challenges in this particular situation. While dimensionality reduction methods may assist in the initial detection of issues, they effectively eliminate data and may weaken a model's ability to make accurate predictions. Nevertheless, because PCA implies that all connections are linear, it may lead to a simplified depiction of the nonlinearities that are evident in several predictors of cardiovascular disease. Despite the aforementioned points, it remains uncertain to what extent PCA contributes to the prediction of cardiac events and the balance between eliminating irrelevant variables (phase 1) without compromising the potential predictive capability of models [6].

Random Forest (RF) is a kind of supervised learning 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. Random Forest is a machine learning method that 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. The bagging technique involves the creation of many decision trees that are produced individually and then merged to provide a final result. This is yet another remarkable machine learning algorithm.

The algorithm randomly selects features and then divides the node based on the feature that will be responsible for constructing particular splits in the mode creation process. The effect is further enhanced by including random breakpoints into each feature. The data is analysed using Random Forest to assign scores to each of the characteristics. 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 Decision Tree (DT) approach is very adaptable and extensively accepted due to its simplicity, which is considered one of its key strengths. A perennial plant with a single woody stem, branches, and leaves, are typically reaching a significant height. The system consists of three nodes: the chance node, the choice node, and the end node. The chance node represents the anticipated result of a certain function, whereas the decision node indicates the possibility of two outcomes after invoking the service until another call is made. End Node: The ultimate node that concludes every trip with a conclusive outcome Request Body refers to the content or data that is included in a request. A decision tree originates with a node known as the root, which then divides into several

branches or nodes. The master node is partitioned using a Random forest classifier. Each node in the tree contains data-specific information on the optimal choice to make at a particular testing stage, while each link represents a decision rule associated with the nodes. There are two methods for displaying or imaging the tree: using the rule and Gini index, and using the entropy-rule as a criterion. It is the most straightforward and transparent forecasting models.

Cleaning and normalizing the raw data is performed early on in the model creation process to guarantee uniformity and efficiency during training and testing. Data normalization is essential for the best possible performance of many different machine learning algorithms since it standardizes the data and brings all features to a consistent scale. Model convergence is improved, larger scale features are not allowed to dominate,

The core premise of deep learning is represented by this duality: training the model on known data and testing its results on unknown data. To ensure the model's validity and utility in real-world settings, the will train and test it on separate datasets. This will allow us to create a model that not only learns from the training set but also generalizes effectively to new, unknown data. Building reliable deep learning models relies heavily on a thorough data preprocessing strategy, with normalization and smart data splitting being two of the most important components.

4) Classification and prediction

Following PCA-based feature selection, the dataset is partitioned into a training set & a testing set with an 80:20 split. The proposed model is trained with data from the training set. The prediction model uses the testing set to classify and predict HD. The classification is the last but not the least step for implementing the new HD prediction model, where several classifiers are involved in classifying coronary heart disease. This work uses four machine learning classifiers to classify heart attacks with hyperparameter settings to predict HD in a patient. These classifiers are:

- AdaBoost classifier [23-24]
- Gradient Boosting classifier [25]
- Extra tree classifier [26]
- CatBoost classifier [27]

In the field of cardiac disease prediction, the choice of classifiers is essential in achieving accurate and dependable results. In situations when both predicted accuracy and interpretability are crucial, classifiers such as AdaBoost, Gradient Boosting, Extra Trees, and CatBoost are used because of their distinct advantages in handling classification tasks.

Once the classification is done, the prediction is made by calculating the confusion matrix, which helps predict heart disease. Also, several parameters are calculated to measure the model performance and provide the early detection of heart disease by getting more precise classification and prediction results. The AdaBoost classifier best provides a more accurate prediction, as shown in the results section.

Making Predictions with AdaBoost:

and training is more stable and effective as a whole as a result of normalization.

The intentional separation of the dataset into training and testing groups, however, is an essential part of the machine learning pipeline. Standard procedure requires that 80% of the cleaned and normalized data be used for training the model, while the remaining 20% is to be used for testing and assessing its efficacy. The chosen proportion is meant to provide a happy medium between giving the model enough information to learn patterns and giving a fair assessment of its generalization abilities.

It would be impossible to emphasize the importance of this method of data division. It allows the model to pick up information from the training dataset while also being evaluated on data it has never seen before (the test dataset).

Determining the weighted sum of weak classifiers enables one to make predictions. Every weak learner will determine a projected value for an incoming feature instance to be either1.0 or -1.0, depending on which value is greater. The step score of each poor learner is factored into the calculation of the estimated values. The estimate for the ensemble classifier is obtained by adding together all of the predictions that have been weighted. If the total is a positive number, then its first class is the one that can be predicted; if the total is a negative number, then its second class can be predicted. This classifier works in the following steps:

Algorithm 1: AdaBoost classifier algorithm

Input: Heart disease dataset

Output: Classification results

Steps:

Step 1. At the start, AdaBoost would select a training sample through the random selection process.

Step 2. AdaBoost model is trained iteratively by selecting a training dataset depending upon the accuracy of such a prediction using the most recent training.

Step 3. It gives the observations that were incorrectly categorized a greater amount of weight to ensure that, in the subsequent iteration, those observations would be given a high likelihood of categorization.

Step 4. In addition, it determines how much weight to give the trained classifier for each iteration by evaluating the classification accuracy of the classifier. A classifier that offers more accurate results will be provided more weight.

Step 5. This method would continue to execute until the entire training data is fitted with no errors or the maximum number of set estimators has been met.

Step 6. The need to "vote" across all the developed learning algorithms to categorize.

More Detailed Explanation of Adaboost

Boosting is an example of an ensemble method that employs multiple weak learners to improve on a single set of data. Classification performance can be enhanced through iterative boosting. Freund and Schapire's Adaptive Boosting (AdaBoost) algorithm is the most well-known boosting

technique. One of the most widely used and studied boosting algorithms, AdaBoost, has found widespread use and has been extensively investigated because it was the first practical boosting algorithm.

The AdaBoost technique 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 labelled samples' expected labels. The resulting error function is used to assign a weight to each spectrum in the subsequent iteration.

The result of this is that misclassified spectra are given more weight, and correctly classified ones are given less. The spectra that haven't been properly classified will then be the subject of subsequent iterations. Every iteration's is weighted weak learners are given access to the unseen data in order to establish the projected class of a previously unseen test subject. A weighted majority vote is then conducted to decide the winner [26].

The proposed algorithm for the newly developed prediction model is given below in algorithm 2.

Algorithm2. Proposed prediction model

Input: Heart disease Cleveland data

Method:

Step 1. Begin
Step 2. Heart disease data gathering from the Kaggle repository
Step 3. Prepare the data by a preprocessing mechanism
Step 4. Null value check
Step 5. Standard scaler
Step 6. Dimensionality reduction using the Random Forest, Decision Trees and PCA method and selecting essential features
Step 7. Dataset separation into two sets
Step 8. Training set (80%)
Step 9. Testing set (20%)
Step 10. Use four machine learning classifiers with hyperparameters settings to training
Step 11. AdaBoost classifier
Step 12. Gradient Boosting classifier
Step 13. extra tree classifier
Step 14. CatBoost classifier
Step 15. Calculate classification results in terms of performance metrics
Step 16. Test the models
Step 17. Provide prediction results through confusion matrices
Step 18. Stop

Output: Heart disease prediction (Yes or No)

IV. RESULTS AND DISCUSSION

The proposed boosting-based predictive model is experimented with using Python programming in the Jupyter Notebook IDE. Several Python libraries like pandas, NumPy, and matplotlib have been used. There are several performance measurement parameters to assess the proposed predictive model's performance.

This section displays results in EDA, tabular and graphical representations. Lastly, experimentation has compared the proposed predictive model with advanced ML models. These experiments have been implemented on the HD dataset of the Kaggle repository, and details are given below.

V. EXPERIMENTAL ENVIRONMENT

- **Required hardware:** The HP workstation is equipped with Windows 10, a 1TB HD, an i7 CPU, 32GB of RAM, and other hardware-specific technical tools. Installing simulation tools and making use of software technologies are both encouraged by the proposed method.
- **Required Software Technologies:** Python, a computer language, and the Jupyter Notebook, a software environment, were used in this study.

A. Dataset descriptions

This research used a subset of thirteen features [23] to create a technique related to clinical circumstances. Clinical variables considered related "SEX," "AGE," "TRESTBPS," and "CP;" "routine test data FBS," "RESTECG," & "CHOL;" exercise electrocardiography test with features "EXANG," "THALACH," "OLDPEAK," & "SLOPE;" & non-invasive test "CA," & "THAL" as shown Table II. Furthermore, the label was NUM.

TABLE II FEATURES DETAILS OF HD DATASET.

Features	Value
AGE	Age (in years)
SEX	1 indicates male; 0 indicates female
Chest Pain (CP) type	1 indicates typical angina; 2 indicates atypical angina; 3 indicates non-angina pain; and 4 indicates asymptomatic
Treetops	Systolic blood pressure at rest (in mm Hg 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 of > 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
exam	Exercise-induced angina (1 indicates yes; 0 indicates no)
Halacha	Maximum heart rate achieved
Slope (slope of peak exercise ST segment)	1 indicates upsloping; 2 indicates flat; 3 indicates downloading
old peak	Exercise-induced ST depression relative to rest
target	Yes=1 or No=0
thal	Exercise Thallium heart scan: 3 indicates normal; 6 indicates fixed defect; 7 indicates a reversible defect
Ca	No. of major vessels (0 to 3) colored by fluoroscopy
AGE	Age (in years)

B. Hyperparameters setting

These hyperparameters were chosen through experimentation and optimization to achieve the best possible performance for each specific machine learning classifier in the given context.

1) AdaBoost Classifier:

- **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 decision trees). This ensemble employs 98 decision trees, as the parameter value 98 indicates.

2) Gradient Boosting (GB) Classifier:

- **n_estimators:** GB also creates an ensemble by combining several decision trees. One hundred decision trees are used in this case.
- **max_depth:** Each decision tree's maximum depth is set to one via the max depth option, making them shallow. This prevents overfitting and assures 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.

Extra Trees Classifier:

- **n_estimators:** Extra Trees, often called Extremely Randomised Trees, is an ensemble method that uses a number of randomised decision trees. The ensemble is built using 6 decision trees in this case.
- **random_state:** Setting this to 98, like AdaBoost, ensures reproducibility by seeding the random number with a specific seed.

3) CatBoost Classifier:

- **Iterations:** Like AdaBoost and GB, CatBoost builds an ensemble from independent decision trees. In this study, perform 5 iterations.
- **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. Novelty of the model

The study employs established techniques like Random Forest, Decision Trees and PCA for feature reduction and commonly used machine learning classifiers such as Adaboost, Gradient Boosting, Extra Trees, and CatBoost. However, it introduces several novel aspects worth highlighting.

Firstly, 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.

Secondly, the application of Random Forest, Decision Trees and PCA for CHD prediction is optimized through parameter fine-tuning and exploring relevant dimensionality reduction techniques. These customizations ensure the model's effectiveness in the specific domain.

Thirdly, selecting the four machine learning classifiers is driven by carefully considering their suitability for CHD prediction. Customizations are made to enhance their performance within the CHD context, contributing to the novelty of the approach.

Furthermore, the study uncovers unique insights and patterns in CHD prediction that may not have been explored in existing literature. These include specific feature importance rankings, model interpretability, and unexpected correlations, providing novel contributions to the field.

Lastly, the research 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. This real-world impact emphasizes the importance and novelty of the study's findings.

In this study, you have employed four machine learning classifiers: Adaboost, Gradient Boosting, Extra Trees, and CatBoost. Each of these models plays a specific role in CHD prediction task. Let's discuss the roles, reasons for their use, and whether they were used individually or as an ensemble.

D. Role of the proposed model

1) Adaboost (Adaptive Boosting):

- **Role**: Adaboost is a strategy for increasing the performance of weak learners like decision trees by giving weights to data points and iteratively changing those weights to focus on misclassified cases.
- **Reason for Use:** Adaboost was chosen because of its ability to improve the accuracy of weak classifiers, making it suitable for improving the predictive power of basic models like decision trees.
- **Usage**: Adaboost can be used alone or in conjunction with others. It is generally used to improve the performance of other classifiers; however, it appears to be employed independently in this case.

2) Gradient Boosting:

- **Role:** Gradient Boosting builds an ensemble of decision trees successively, with each tree correcting the flaws of the one before it.
- **Reason for Use:** Gradient boosting is well known for its great predicted accuracy and ability to handle complex data interactions. It is picked to model the CHD prediction issue adequately.
- **Usage**: Gradient Boosting can be employed singly or as an ensemble. It is frequently used as an independent model due to its high predictive ability.

3) Extra Trees:

- **Role**: An ensemble learning technique based on decision trees is called Extra Trees, commonly called Extremely Randomized Trees. It incorporates randomness into tree construction to minimize overfitting.
- **Reason for Use**: Extra Trees are utilized because of their resistance to overfitting and ability to capture significant features in data while lowering variation.
- **Usage:** Individual and ensemble use of Extra Trees. Many utilize it as an individual model to benefit from its noise-reduction properties

4) CatBoost:

- **Role**: CatBoost is an effective gradient-boosting method for working with categorical features. Categorical data encoding is handled manually.
- **Reason for Use**: CatBoost is selected due to its ability to perform well with a mixture of categorical and numeric data, frequently encountered in healthcare datasets such as CHD prediction.
- **Usage:** CatBoost is usually used independently, benefiting from its category feature processing.

In the study, these machine-learning models were used individually rather than as an ensemble. Each model serves a specific purpose, and their performances are evaluated using metrics such as Accuracy, Precision, Recall, and F1-score. This approach allows you to assess the strengths and weaknesses of each model in the context of CHD prediction and make informed decisions about model selection and hyperparameter tuning.

E. Performance Measurement Parameters

There are some parameters for measuring the performance of the proposed system. These parameters are essential to calculate because they are needed to 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, which are briefly defined.

1) Confusion Matrix (Cm)

CM is a table-based representation of ground-truth labeling vs. predicted results. Each row of CM signifies the cases in the predicted class, whereas every column represents cases inside an actual class. It is not technically a performance statistic but a foundation for other measures to analyze the findings.

We should first make a presumption about the null hypothesis to understand CM. For instance, let us say that the Null Hypothesis H0 is "The person has heart disease".

		Predicted	
		Heart attack (Yes)	**Heart attack (No)**
Ground truth	**Heart attack (Yes)**	TP	FP
	Heart attack (No)	FN	TN

Fig. 2. Confusion Matrix for H0.

In the confusion matrix, each row contains an assessment parameter. Let us look at these aspects each one individually:

- TP (True Positive) denotes how several positive class instances the model properly predicts [28].
- FP (False Positive) The number of false positives represents how several negative class instances the model can predict wrongly. In statistics terms, this factor reflects Type-I error. H0 determines the position of the error in CM.

FN (False Negative) represents no. the positive class instances that the model predicts wrongly [7]. In statistics terms, this component reflects Type II error. This incorrect location in the confusion matrix is also affected by the H0 chosen, as shown in Figure 2.

2) Classification Metrics

- **Accuracy:** Classification accuracy is likely the easiest statistic to use & execute, which is determined as the of true predictions separated by total predictions multiplied by a hundred [28].
- The ratio of TPs to overall expected positives is known as precision.

$$Precision = \frac{TP}{TP+FP} = \frac{Correctly\ identifies\ Heart\ patients}{\substack{Correctly\ identifies\ Heart\ patients+ \\ incorrectly\ labelled\ heart\ patients\ as\ heart\ attack}} \quad (2)$$

Precision: The precision metric is concerned with Type-I errors (False Positive). When reject a correct H0, then make a Type-I error. As a result, Type-I error wrongly categorizes heart patients as having no heart attack [7,29].

Recall: The percentage of actual positives across all positives in the ground truth.

$$Recall = \frac{TP}{TP+FN} = \frac{Correctly\ identifies\ Heart\ patients}{Correctly\ identifies\ Heart\ patients} \quad (3)$$

Recall

The recall measure is concerned with type II errors (False Negative). If accepted a false H0, then commit a type-II error. In this situation, the type-II error involves mislabeling non-heart attack patients as having a heart attack.

F1-score: Recall and precision are utilized in the F1-score measure. It is a harmonic mean of two. The basic equation for 2 is:

$$F1 - score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \quad (4)$$

This study utilizes the F1 score because it demonstrates high precision and recall. It has a high recall-to-precision ratio and performs well on unbalanced classification problems. The F0.5 or F2 score is not used because it depends on the classification problem's objectives and priorities.

1. **F0.5 score:** The F0.5 score emphasizes precision over recall by giving precision greater weight. This might be chosen when the cost of false positives is significantly higher than that of false negatives. Authors might use F0.5 when they want to prioritize minimizing false positives while still considering recall.

2. **F2 Score:** The F2 score, on the other hand, places more emphasis on recall than precision by giving more weight to recall. This could be chosen when the cost of false negatives is significantly higher than that of false positives. Authors might use F2 when they want to prioritize identifying as many true positives as possible while accepting some false positives.

AUC-ROC Score: It is more usually recognized. It uses TPR and FPR.

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN} \quad (5)$$

- Generally speaking, TPR or recall represents a portion of positive samples correctly labelled as positive compared to all positive samples. Alternatively, we would discard a greater TPR and fewer positive samples.
- FPR or fallout generally equates to a ratio of the negative sample erroneously interpreted positively when contrasted to all negative samples. Alternatively, the more negative samples mistakenly classify as positive, the higher the FPR.

Precision-Recall curve: This curve depicts trade-offs across recall and precision for various threshold values. A high AUC suggests both strong recall & high precision, with low FPR relating to high precision and low FNR relating to high recall [7].

F. Results

This section displayed different computed results in data visualization, the number of components for feature extraction, and classification results.

1) Exploratory data analysis visualization

Different visualization representations display several features in the HD dataset. The process of developing plots & other visuals when coping with somewhat new geographical information is termed exploratory visualization. These plots frequently serve a specific purpose & serve as an aid in an expert's endeavor to handle (geographic) issues [29]. Here, graphics visuals like heatmap plots, Correlation matrix, Bar plots, Pie plots, replot, Catplot, and histogram plots are considered.

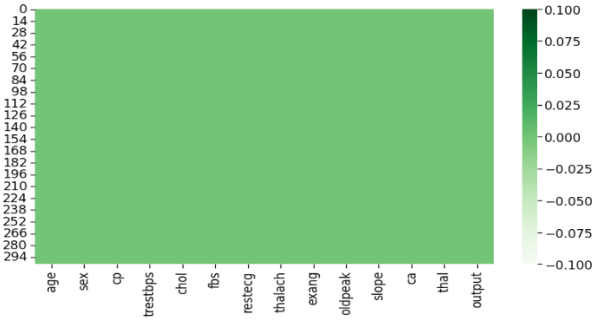


Fig. 3. Heatmap for null values.

Figure 3 shows the heatmap plot of the heart disease data set to display the null values, which present null values for each of the features. It consists of a scale labeled -0.1 to 0.1 to display the level of the presence of the null value [30-31].

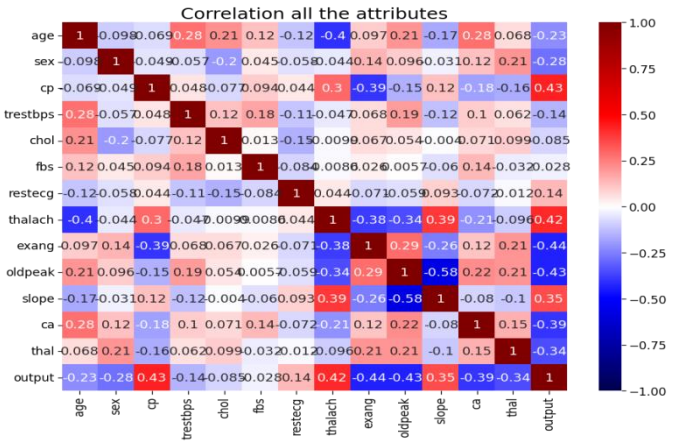


Fig. 4. Correlation Matrix.

Figure 4 displays a correlation matrix in which every feature is correlated. Its cell value is shown from minus one to plus one through different color levels. It is displayed with the feature brown color with the highest correlation and dark blue color indicating the lowest correlation. Diagonally, each attribute has the highest correlation value. Apart from this, the highest correlation is between the chol and that features value having 0.099, and the least is between output and exchange feature value having -0.44.

This study examines how Machine Learning (ML) algorithms may detect Coronary Heart Disease (CHD) early. The goal is to evaluate the performance of three ML algorithms: Random Forest (RF), Logistic Regression (LR), and Support Vector Machine.

The heart disease dataset seems to have been a well-rounded fusion of various sources, contributing to its broad range and absolute magnitude. This dataset included 281 female and 909 male patients' age, gender, chest pain type (4 values), resting blood pressure (mmHg w/in circulation), and serum cholesterol (mg/dl as in heart attack) (see data tab for details).

The investigation used hyper-parameter optimisation and repeated cross-validation to provide robust model performance. Accuracy, Specificity, Sensitivity, F1-Score, PPV, and NPV were used to evaluate performance. With 0.929 accuracy, RF outperformed SVM and LR.

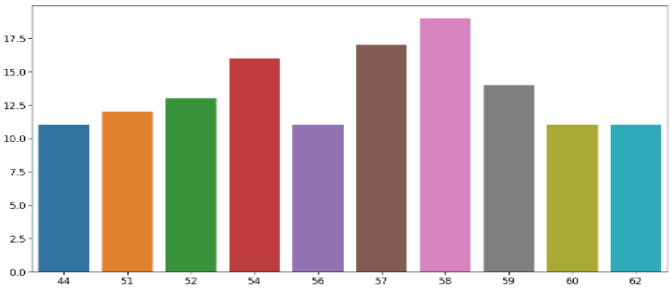


Fig. 5. Bar graph of age counts.

Figure 5 bar graph 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. The graph positions the age categories based on their respective counts, with age 44 having the lowest (0.5) and age 58 having the highest.

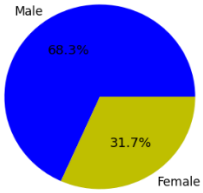


Fig. 6. Pie plotting for % of sex counts.

Figure 6 pie chart shows the percentage of sex counts for a group of people. The chart has two sections, one for males and one for females. The male section is blue and takes up 68.3% of the chart, while the female section is yellow and takes up 31.7% of the chart.

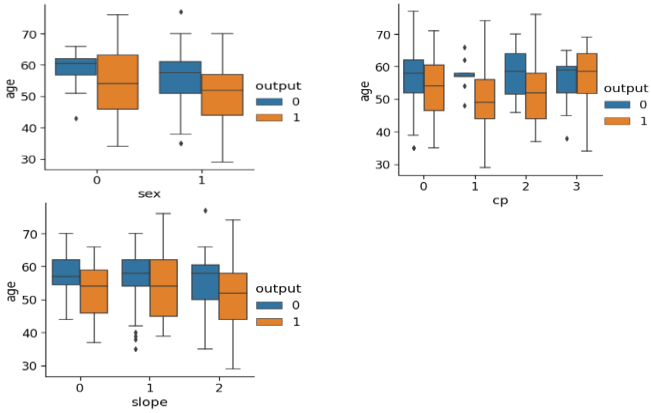


Fig.7. Box plots of various attributes.

Figure 7 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.

The box plots show the median, quartiles, and range of age for each category and output value. The outliers are marked as black dots.

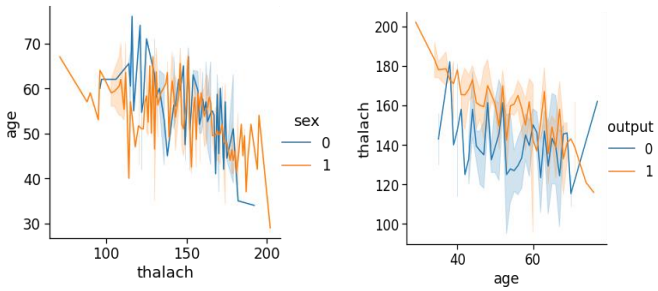


Fig. 8. scatter, plot between age thalach, sex, and output.

Figure 8 shows two scatter plots that plot the relationship between age, thalach, sex, and output. The graph on the left shows how thalach varies with age for different sex values. The graph on the right shows how thalach varies with age for different output values. The lines are colored orange and blue, indicating sex 0 and 1 on the left graph and output 0 and 1.

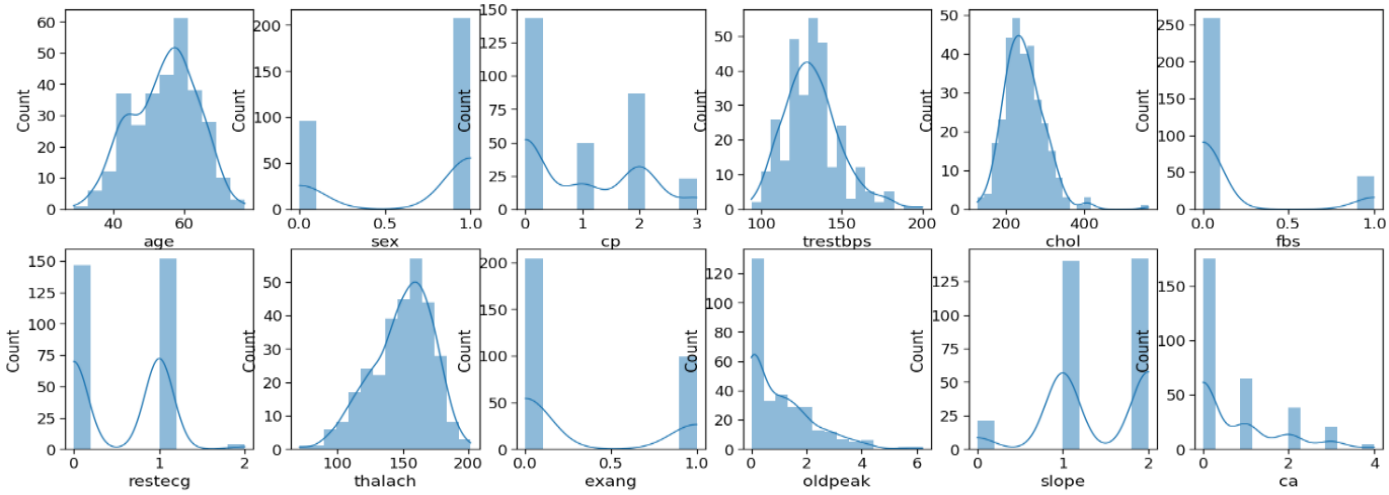


Fig. 9. Histogram of heart disease dataset.

Figure 9 shows a set of histograms and density plots that visualize the distribution of different variables in a heart disease dataset. The image has 14 plots, each showing one variable. The variables are restecg, trestbps, ca, exang, age, chol, fbs, oldpeak, thalach, slope. The plots are blue in color and have a white background. The x-axis of each plot shows a range of values for the variable, & 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.

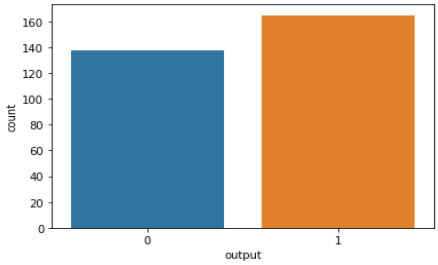


Fig. 10. Data labelling count of heart disease outcome.

Figure 10 depicts the labelling 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.

2) Feature selection results

Several features in data reduce model learning speed; therefore, PCA is utilized to minimize data dimensions to build a model fast with less machine effort and help increase the proposed model efficiency [32].

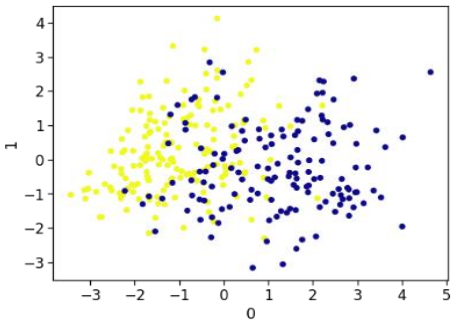
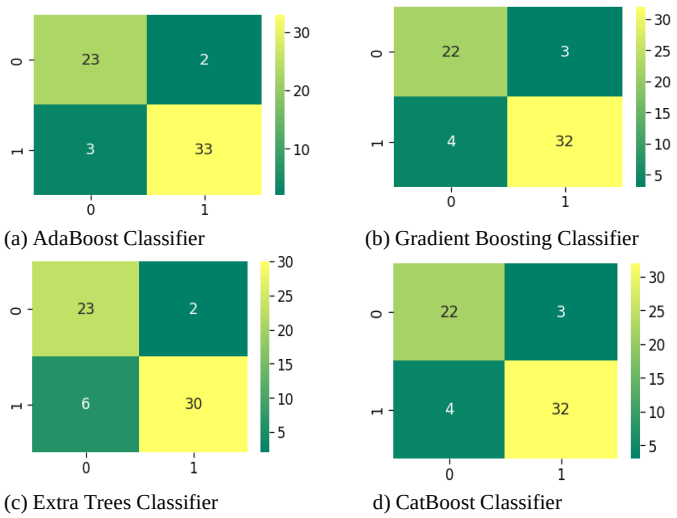


Fig. 11. Features selection using PCA.

Figure11 depicts the extracted features after applying the PCA extraction method for 14 different features. After PCA, new d29 has been reduced to 2 features, with the same no. And two target components, Yes (1) or No (0), of rows as an original feature.

3) Confusion matrices

Multiple confusion matrices are displayed for all four machine-learning classifiers. These are binary confusion matrices for telling us whether a person has heart disease (1) or not (0).



(a) AdaBoost Classifier (b) Gradient Boosting Classifier

(c) Extra Trees Classifier d) CatBoost Classifier

Fig. 12. Testing confusion matrices for different classifiers.

Figure 12 depicts the confusion matrix plots for different classifiers after applying them with different hyperparameter settings. These matrices are plotted between actual and predicted labels to display whether heart attack or not in patients. Figure 12 (a) depicts the testing confusion matrix for the AdaBoost classifier in which correctly positive classified (TP) instances are 33, misclassified instances are 3 (for FN) and 2 (for FP), and correctly negative classified instances (TN) is 23. Figure 12 (b) depicts the testing binary confusion matrix for the Gradient Boosting classifier in which correctly positive classified (TP) instances are 32, misclassified instances are 4 (for FN), and 3 (for FP) and correctly negative classified instances (TN) are 22. Figure 12 (c) depicts the testing binary confusion matrix for the Extra Trees. Classifier in which correctly positive classified (TP) instances are 30, misclassified instances are 6 (for FN) and 2 (for FP), and correctly negative classified instances (TN) are 23. Figure 12 (d) depicts the testing binary confusion matrix for the CatBoost classifier in which correctly positive classified (TP) instances are 32, misclassified instances are 4 (for FN), 5 (for FP), and correctly negative classified instances (TN) are 20 [33-34].

4) Classification results

The classification results for all four machine learning classifiers. These classification reports display class-wise performance results regarding Precision, recall, F1-Score, weighted average, macro average, and accuracy.

	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

(a) AdaBoost Classifier

	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

(b) Gradient Boosting Classifier

	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

(c) Extra Trees Classifier

	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

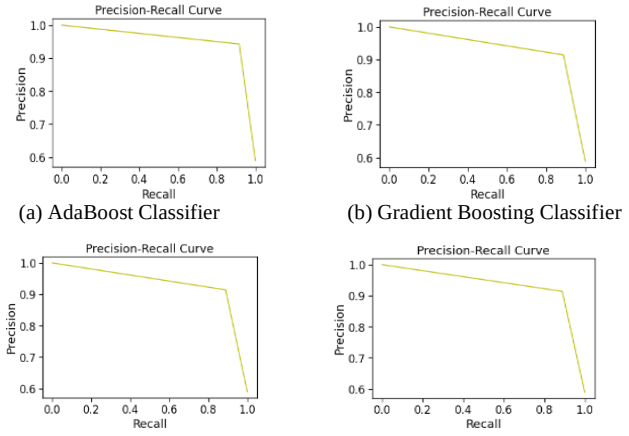
(d) CatBoost Classifier

Fig. 13. Classification results for different classifiers.

After implementing and executing, Figure 13 depicts the classification results for AdaBoost, Gradient Boosting, Extra Trees, and CatBoost classifiers. Figure 13 (a) depicts the classification results for the AdaBoost classifier: class 0 has 97% precision, 92% recall, and 90% f1-score, whereas class 1 has 94% precision and 83% recall, 88% f1-score while accuracy is 87%. Figure 13 (b) depicts the classification results for the Gradient Boosting classifier: class 0 has 85% precision, 88% recall, and 86% f1-score, whereas class 1 has 91% precision, 89% recall, 90% f1-score while accuracy is 87%. Figure 13 (c) depicts the classification results for the Extra Trees classifier in which class 0 has 79% precision, 92% recall, and 85% f1-score, whereas class 1 has 94% precision, 83% recall, 88% f1-score while accuracy is 89%. Figure 13 (d) depicts the classification results for the CatBoost classifier: class 0 has 83% precision, 80% recall, and 82% f1-score, whereas class 1 has 86% precision and 89% recall, 88% f1-score while accuracy is 85%.

5) Curve analysis

This displays the curve analysis for different models for the precision-recall and ROC curves. The precision-recall curve is plotted between precision and recall, whereas the ROC curve analysis is plotted between TPR and FPR [35-36].



(a) AdaBoost Classifier (b) Gradient Boosting Classifier

(c) Extra Trees Classifier (d) CatBoost Classifier

Fig.14. Precision-recall curve plots for multiple classifiers of heart disease.

Figure 14 depicts Precision-recall curve plots for four different classifiers of the HD dataset. Each of the classifiers has displayed individual curve results. This figure shows AdaBoost, gradient boosting, and Extra Trees classifiers have achieved more than 0.9, but the CatBoost classifier achieved less than 0.9 values.

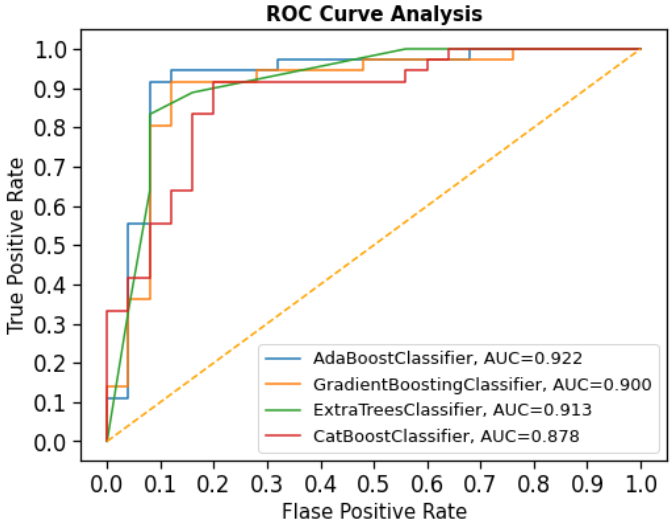


Fig. 15. ROC curve for multiple classifiers of heart disease.

Figure 15 shows the ROC curve analysis, containing the TPR and FPR for all four classifiers. In this curve plot, each classifier is represented by different colors. The CatBoost classifiers have achieved an AUC value of 0.883, which is very low. The AUC values of the remaining three classifiers are greater than 0.962.

This graph illustrates that the AdaBoost (representing blue color) model is the most accurate predictive classifier (AUC = 0.962) for identifying heart attacks.

TABLE III COMPARISON OF DIFFERENT ML CLASSIFIERS.

ML classifiers	Accuracy	F1-score	Precision	Recall	AUC
AdaBoost Classifier	97.803	95.826	95.899	99.803	0.962
GradientBoosting Classifier	96.525	88.556	88.636	88.525	0.930
ExtraTrees Classifier [26]	93.885	96.985	87.832	96.885	0.883
CatBoost Classifier [27]	92.246	95.196	95.194	95.246	0.938
Random Forest Classifier [22]	95.246	95.214	95.228	95.246	0.951
Logistic Regression [22]	95.246	95.214	92.228	93.246	0.938

Table III represents a comparison of different ML classifiers to show performance results. It shows results in terms of different performance metrics.

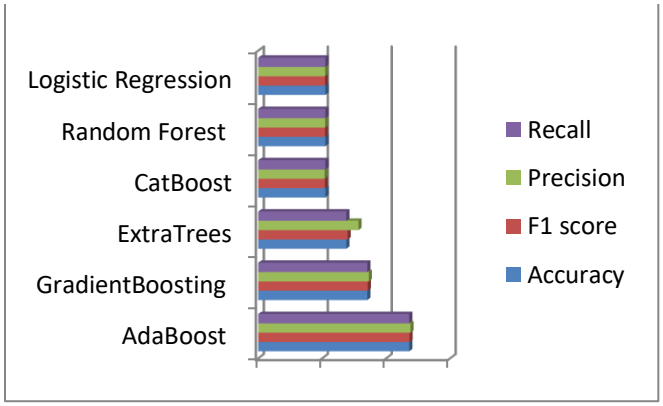


Fig. 16. Comparison bar graph of different classification performance results.

Figure 16 shows displaying the results of several classifications. While the graphic doesn't provide the exact classification problem or dataset, it does show that Logistic Regression, Random Forest, CatBoost, ExtraTrees, GradientBoosting, and AdaBoost are among the classifiers being compared.

Recall, accuracy, precision, and F1 score are the performance parameters that are being compared. What term "recall" is the percentage of accurately detected true positive? How many false positives were truly correct is known as precision. A harmonic mean of accuracy and recall is the F1 score. The total percentage of right categorizations is called accuracy.

Based on the data shown in the bar graph, CatBoost outperformed GradientBoosting and ExtraTrees in terms of F1 score, coming in at about 90%. Logistic Regression's F1 score was the lowest.

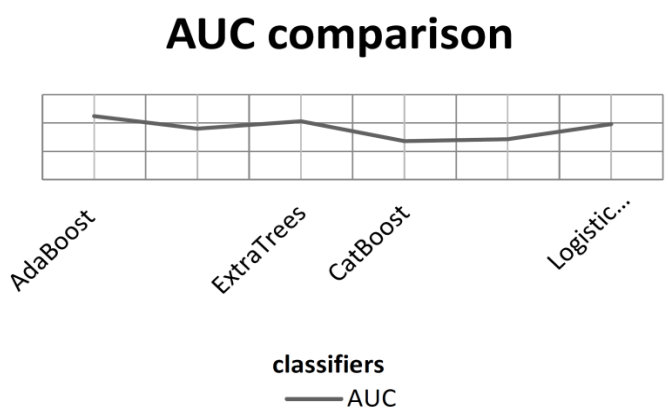


Fig.17. Comparison line graph of curve analysis results.

Figure 17 shows the comparative line graph among different classifiers for measuring the AUC results. Here, an x-axis indicates different ML classifiers & y-axis indicates an AUC value. This comparison shows that the AUC of Gradient Boosting, extra tree, and logistic regression classifiers achieved 93%. However, it was satisfactory, but AdaBoost classifier results are higher than those of three machine learning classifiers by achieving 96.2% AUC result. CatBoost achieved worse AUC results than the other five classifiers, with 87.8% AUC results [37].

VI. DISCUSSION

This study's discussion section emphasizes several significant findings and ramifications associated with using ML techniques for forecasting cardiovascular diseases. Initially, it emphasizes the importance of prognosis in medical decision-making and the need for accurate risk prediction models. Given their high frequency and financial cost, the study's emphasis on heart diseases is justified because it aligns with current healthcare goals. The performance of four different classifiers is 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. Based on this discovery, probably, AdaBoost could diagnose heart disease in real-time and thus has major therapeutic implications. Finally, it can also been seen that the proposed AdaBoost model outperforms other models in case of AUC value and accuracy. The assessment of the difference in the findings of the foregoing employed and original studies emphasises the applicability and relevance of applying the AdaBoost method for predicting heart disease in the healthcare context. What is important to admit, however, is that more research and evidence is needed to identify. Different patients' health requires evaluating the models' performance and transferability to different patient populations for the healthcare applications. Also, while the study mentions the data preprocessing and the feature selection approaches that were implemented and the primary aim of predicting the chances for developing coronary heart disease (CHD), more information on the features of the data set, including the sample size, and data collection would enhance the study's credibility.

- Analysis:
  - Conclusion: The paper provides a highly efficient solution for CAD identification at the prediagnosis stage, as well as a/R method for the selection of the vital characteristics and classification of the female patients' data. This method has suggested one way of enhancing patients' outcomes and management in LBP instances. CAD is well recognized as being one of the most prevalent and fatal conditions affecting people's lives globally.
  - Early diagnosis might mean increased treatment and improved patient outcome. Purpose: The goal of this project is to create a machine learning algorithm that can promptly and efficiently diagnose CAD using minimal possible features.
  - The functional aspect of this above said technique is defined under speed, accuracy and the capacity that can reduce the existence of few characteristics out of thousands which may be potentially useful in clinical field. It responds to the requirements of data integration, the presence of bias in the algorithms and clinical application of the models that are needed.

VII. CONCLUSION

Disease prediction is very important in medical decision-making, and physicians must have a comprehensive understanding of the risks associated with various diseases, this task is accelerated by logical and methodological approaches such as ML approach the application of the. Given the increasing incidence of cardiovascular disease and its significant economic impact on society, healthcare professionals are constantly looking for more effective ways to predict, diagnose, and treat cardiovascular disease. ROC-AUC and precision-recall curve graphs are used to specifically assess the performance of the classifiers on the test data set. The findings show that the AdaBoost machine learning classifier outperforms its competitors in cardiovascular disease detection, with an accuracy of 97.803%, an F1-score of 95.826%, and an accuracy of 0.962%. , the AUC was 94.2%. Finally, the AdaBoost model exhibits robust performance and has the ability to accurately predict cardiac complications in real time. However, further research and validation is needed to determine its effectiveness in various health care settings.

A. Limitations:

The study repeatedly finds difficulties in integrating large amounts of data, potential biases, and the need for extensive clinical validation of each page Further research is needed to address this knowledge gap , have used extensive data collection to verify the validity of the model, ensuring universal compliance with acceptable ethical standards.

B. Future work

The potential of device getting to know in predicting coronary heart infection could be very promising. One way to enhance is via growing extra complicated prediction models, which would possibly encompass methods like deep studying, and through including diverse statistics sources like as genomes and affected person histories. Personalised

medication, which customises remedy techniques primarily based on man or woman chance factors, genetics, and way of life, is an exciting possibility. Implementing real-time tracking structures has the capability to find out problems at an early stage and permit for set off actions, ensuing in reduced healthcare charges.

The integration of machine mastering into medical practice might be facilitated through the usage of standardised information formats and interoperable healthcare structures. In addition to technical progress, it will be critical to tackle moral and regulatory concerns in order to shield affected person privateness, assure statistics protection, and promote an appropriate use of AI inside the healthcare area. By using person-friendly interfaces and cell apps, affected person involvement has the potential to enhance information collection and adherence to treatment programmers, ultimately improving the accuracy of prediction models.

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), 2021, pp 232-237. Doi: https://doi.org/10.1109/ICSECS52883.2021.00049.

[2] Y. Khourdifi, Hassan 1st University, M. Bahaj, and Hassan 1st University, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," Int. J. Intell. Eng. Syst., vol. 12, no. 1, pp. 242–252, 2019. Doi: 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," J. Biomed. Inform., vol. 58, pp. S171–S182, 2015. Doi: https://doi.org/10.1016/j.jbi.2015.09.006.

[4] H. Kim, M. Ishag, M. Piao, T. Kwon, and K. Ryu, "A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries," Symmetry, vol. 8, no. 6, p. 47, 2016. Doi: https://doi.org/10.3390/sym8060047.

[5] 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, 2016. Doi: https://doi.org/10.5120/ijca2016909231.

[6] 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, no. 100330, p. 100330, 2020. Doi: https://doi.org/10.1016/j.imu.2020.100330.

[7] 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), IEEE, pp 92-99, 2018. Doi: https://doi.org/10.1109/SKG.2018.00020.

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

[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, pp. 1–17, 2021. Doi: https://doi.org/10.1155/2021/5581806.

[10] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, "A review of feature selection and classification approaches for heart disease prediction," IJITEE (International Journal of Information Technology and Electrical Engineering), vol. 4, no. 3, p. 75, 2021. Doi: https://doi.org/10.22146/ijitee.59193.

[11] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," BMC Bioinformatics, vol. 21, no. 1, pp. 1-18, 2020. https://doi.org/10.1186/s12859-020-03626-y.

[12] S. Usha, S. Kanchana, S, "Predicting Heart Disease Using Feature Selection Techniques Based on Data Driven Approach," Webology, 18 (4), 2021.

[13] 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, p. 19747, 2020.Doi: https://doi.org/10.1038/s41598-020-76635-9.

[14] 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, 2022. Doi: https://doi.org/10.1007/s42044-022-00104-x.

[15] T. K. Sajja and H. K. Kalluri, "A Deep Learning method for prediction of Cardiovascular Disease using Convolutional Neural Network," Rev. d'Intelligence Artif., vol. 34, no. 5, pp. 601–606, 2020. Doi: https://doi.org/10.18280/ria.340510.

[16] 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, p. 022006, 2020. Doi: http://dx.doi.org/10.1088/1757-899X/981/2/022006.

[17] 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. Doi: https://doi.org/10.1109/ACCESS.2020.3010800.

[18] 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), pp 393-401, 2022. Doi: https://doi.org/10.1109/CISES54857.2022.9844406.

[19] 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. Doi: https://doi.org/10.1109/ACCESS.2021.3116974.

[20] 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, pp. 1–11, 2021. Doi: https://doi.org/10.1155/2021/8387680.

[21] 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, pp. 9-16, 2012. Doi: https://doi.org/10.1109/ICDMW.2012.29.

[22] S. Prusty, S. Patnaik, and S. Kumar Dash, "Comparative analysis and prediction of coronary heart disease," Indonesian Journal of Electrical Engineering and Computer Science, 27 (2), 944-953, 2022. Doi: https://doi.org/10.11591/ijeecs.v27.i2.pp944-953.

[23] G. Deivendran, S. Vishal Balaji, B. Paramasivan, and S. Vimal (Correspon, "Coronary illness prediction using the AdaBoost algorithm," Sensor Data Analysis and Management. Wiley, pp. 161–172, 26-Nov-2021. Doi: https://doi.org/10.1002/9781119682806.ch10.

[24] 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. Doi: https://doi.org/10.1039/D0AN02155E.

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

[26] R. Shafique, A. Mehmood, S. Ullah, and G. S. Choi, "Cardiovascular disease prediction system using extra trees classifier," , Research Square, 2019. Doi: https://doi.org/10.21203/rs.2.14454/v1.

[27] X. Zhang, M. Wang, W. Wei, Y. Xu, L. Gao, Y. Sun, Z. Ma, S. Wang, "An accurate diagnosis of coronary heart disease by Catboost, with easily accessible data," Journal of Physics: Conference Series, IOP

Publishing, vol. 1955, no. 1, p. 012027, 2021. Doi: https://doi.org/10.1088/1742-6596/1955/1/0.

[28] M. Jalil Jassim 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, pp. 183-191, 2022. DOI: https://doi.org/10.11591/ijeecs.v28.i1.pp183-191.

[29] M.-J. Kraak, "Exploratory visualization," in Encyclopedia of GIS, Shekhar, S., Xiong, H., Eds. Springer US: Boston, MA, 2008, pp. 301–307. Doi: https://doi.org/10.1007/978-0-387-35973-1_397.

[30] C. J. J. Sheela and G. Suganthi, "Morphological edge detection and brain tumor segmentation in Magnetic Resonance (MR) images based on region growing and performance evaluation of modified Fuzzy C-Means (FCM) algorithm," Multimedia Tools and Applications, vol. 79, no. 25–26, pp. 17483–17496, 2020. Doi: https://doi.org/10.1007/s11042-020-08636-9.

[31] 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. DOI: https://doi.org/10.1109/ACCESS.2020.3001149.

[32] 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. Doi: https://doi.org/10.1109/ACCESS.2019.2948266.

[33] M. J. Ghrabat, G. Ma, P. L. P. Avila, M. J. Jassim, 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, 9 (4), 483-489, 2019.

[34] S. M. Alanazi and G. S. M. Khamis, "Optimizing Machine Learning classifiers for enhanced Cardiovascular Disease prediction," Eng. Technol. Appl. Sci. Res., vol. 14, no. 1, pp. 12911–12917, 2024. doi: https://doi.org/10.48084/etasr.6684.

[35] L. Corbat, M. Nauval, J. Henriet, and J.-C. Lapayre, "A fusion method based on Deep Learning and Case-Based Reasoning which improves the resulting medical image segmentations," Expert Systems with Applications, vol. 147, no. 113200, p. 113200, 2020. Doi: https://doi.org/10.1016/j.eswa.2020.113200.

[36] B. Alshawi, "Utilizing GANs for credit card fraud detection: A comparison of supervised learning algorithms," Eng. Technol. Appl. Sci. Res., vol. 13, no. 6, pp. 12264–12270, 2023. doi: https://doi.org/10.48084/etasr.6434.

[37] M. Anam, M. Hussain, M. W. Nadeem, M. Javed Awan, H. G. Goh, S. Qadeer, "Osteoporosis prediction for trabecular bone using machine learning: a review," Computers, Materials & Continua (CMC), 67 (1), 2021. https://doi.org/10.32604/cmc.2021.013159.

[38] V. Shorewala, "Early detection of coronary heart disease using machine learning," Journal of Informatics in Medicine, 26, 2021. https://doi.org/10.1016/j.imu.2021.100655

[39] M. Sebastiani, C. Vacchi, A. Manfredi, G. Cassone, "Personalized Medicine and Machine Learning: A Roadmap for the Future," Journal of Clinical Medicine, 11 (14):, 2022. https://doi.org/10.3390/jcm11144110

[40] F. Özbilgin, Ç. Kurnaz, E. Aydın, "Prediction of Coronary Artery Disease Using Machine Learning Techniques with Iris Analysis," Diagnostics," 13 (6), 2023. https://doi.org/10.3390/diagnostics13061081

[41] C. Eyupoglu, O. Karakuş, "Novel CAD Diagnosis Method Based on Search, PCA, and AdaBoostM1 Techniques," Journal of Clinical Medicine, 13 (10), 2024. https://doi.org/10.3390/jcm13102868

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