computation

MDPI

Article

A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques

Najmu Nissa [1], Sanjay Jamwal [1,*] and Mehdi Neshat [2,3]

[1] Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri 185234, Jammu and Kashmir, India; najmunissa@bgsbu.ac.in
[2] Centre for Artificial Intelligence Research & Optimisation, Torrens University Australia, Brisbane, QLD 4006, Australia; mehdi.neshat@torrens.edu.au
[3] Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, Sydney, NSW 2007, Australia
* Correspondence: sanjayjamwal@bgsbu.ac.in

Abstract: This paper addresses the global surge in heart disease prevalence and its impact on public health, stressing the need for accurate predictive models. The timely identification of individuals at risk of developing cardiovascular ailments is paramount for implementing preventive measures and timely interventions. The World Health Organization (WHO) reports that cardiovascular diseases, responsible for an alarming 17.9 million annual fatalities, constitute a significant 31% of the global mortality rate. The intricate clinical landscape, characterized by inherent variability and a complex interplay of factors, poses challenges for accurately diagnosing the severity of cardiac conditions and predicting their progression. Consequently, early identification emerges as a pivotal factor in the successful treatment of heart-related ailments. This research presents a comprehensive framework for the prediction of cardiovascular diseases, leveraging advanced boosting techniques and machine learning methodologies, including Cat boost, Random Forest, Gradient boosting, Light GBM, and Ada boost. Focusing on "Early Heart Disease Prediction using Boosting Techniques", this paper aims to contribute to the development of robust models capable of reliably forecasting cardiovascular health risks. Model performance is rigorously assessed using a substantial dataset on heart illnesses from the UCI machine learning library. With 26 feature-based numerical and categorical variables, this dataset encompasses 8763 samples collected globally. The empirical findings highlight AdaBoost as the preeminent performer, achieving a notable accuracy of 95% and excelling in metrics such as negative predicted value (0.83), false positive rate (0.04), false negative rate (0.04), and false development rate (0.01). These results underscore AdaBoost's superiority in predictive accuracy and overall performance compared to alternative algorithms, contributing valuable insights to the field of cardiovascular health prediction.

Keywords: heart disease; diagnosis; machine learning; boosting; ensemble learning; comparative framework

Citation: Nissa, N.; Jamwal, S.; Neshat, M. A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques. *Computation* **2024**, *12*, 15. https://doi.org/10.3390/computation12010015

Academic Editors: Yudong Zhang and Francesco Cauteruccio

Received: 13 December 2023
Revised: 10 January 2024
Accepted: 11 January 2024
Published: 16 January 2024

1. Introduction

With 17.9 million deaths per year, cardiovascular disease (CVD) is the leading cause of mortality worldwide [1]. CVD is a broad term that refers to any illness of the cardiovascular system. According to the World Health Organization (WHO), heart disease accounts for 80% of these fatalities [2]. Furthermore, socioeconomic factors such as jobs and money have an influence on the death rate via the impact on risk factors connected to lifestyle before and after cardiac disease [3]. The best way to decrease these deaths is to detect them immediately. The ability to predict the existence of HD is critical for administering the necessary therapy on time. The incidence of HD is predicted to quadruple by the year 2020, and it is expected that in 2050, one person will develop the disease every 30 s [4]. The symptoms and occurrence of heart disease vary according to the lifestyle of humans. It

generally comprises pain in the chest, jaw ache, neck pain, back pain, belly issues, fingers and shoulder pains and significant shortness of breath [5]. The various types of heart disease are shown in Table **??** below [6].

Table 1. Various types of popular heart disease.

S. No	Heart Disease	Description
01	Coronary artery disease	Damage to the heart's major arteries. These include: • Atherosclerosis: the constriction of blood arteries as a result of fat deposition • Primarily cholesterol and arteriosclerosis: the hardening of blood arteries • Caused by calcium and cholesterol deposits
02	Hypertension	The condition in which there is an excessive force of blood against the arterial walls is characterized as being hypertensive.
03	Cardiac Arrest	A state in which the cessation of cardiac contractions, respiratory movements, and loss of consciousness occur is commonly referred to as cardiac arrest.
04	Arrhythmias	An irregular rhythm of the heart, known as bradycardia, which manifests as a decreased heart rate, and tachycardia, which presents as an increased heart rate, are both observed phenomena.
05	Peripheral AD	The narrowing of arteries, resulting in a constriction of blood flow, is classified as a condition known as arteriosclerosis.
06	Ischemia	Restricted blood supply to heart muscles.

As a result, it has become critical for the precise and accurate prognosis of heart-related disorders. Many academics from all around the world began researching the prediction of cardiac problems by analyzing enormous databases for this purpose [7]. Various deep learning approaches have the capability of working on enormous datasets and drawing relevant results [8]. Deep learning models are based on numerous algorithms, and so these algorithms have become significant for properly predicting the presence or absence of cardiac ailments. Different researchers all over the world have worked on heart disease prediction using various techniques of machine learning, deep learning, and fuzzy logic [9,10], but still there are some shortcomings, which are given below:

- Almost all the research work with the target of predicting heart disease using techniques of machine learning, deep learning, and fuzzy logic [11] have been carried out using various parameters, but still, there is inadequate parameter tuning and parameter evaluation.
- Lack of use of different discretization techniques, multiple classifiers, techniques of voting, and other decision tree algorithms (Gini index, Gini ratio).
- The technical issues with regard to overfitting.
- Selection and usage of proper tools, proper pre-processing of datasets, and the use of advanced machine learning algorithms to reduce time complexity should be incorporated.

1.1. Socio-Economic Impact of Heart Diseases

In practically all Western nations, socioeconomic disparities in the frequency and occurrence of CVD fatalities have been documented [1]. When compared to other countries in the same region, India has a much greater prevalence of heart disease. A prevalence rate of 11% is concerning [2] despite the fact that India has a very large population. This is due to the fact that India has a very high population density. The prevalence rate sheds light on emerging patterns of sickness occurrence. It is estimated that cardiovascular disease, stroke, and diabetes cost 1% of world GDP, which is equivalent to USD 236.6 billion, between the years 2005 and 2015 [3]. It was projected in the year 2000 that adult Indians had lost a total of 9.2 billion years of productive life, which contributed to the overall economic loss. According to research conducted in 2015, the average lifespan of an Indian man is

estimated to be 67.3 years, whilst the average lifespan of an Indian woman is estimated to be 69.9 years [4]. It was projected in the year 2000 that adult Indians had lost a total of 9.2 billion years of productive life, which contributed to the overall economic loss. As a result, India contributes to the growing population of people who are affected by heart disease as a result of the increased number of elderly people who are prone to developing the condition. An essential medications list (EML) is a project that the World Health Organization (WHO) is working on to enhance healthcare systems in countries with low and intermediate incomes.

Age, modifications in lifestyle and diet, and rapidly expanding social issues, such as hospitalization, all have an influence on the likelihood of acquiring heart-related disorders and the advancement of preexisting conditions. Research indicates that heart disease is responsible for 80% of fatalities and 85% of disabilities in countries with poor and intermediate incomes [5]. The forecasts of total deaths attributable to cardiovascular disease that are shown in Table **??** include the period between the years 2000 and 2030. The treatment of expensive illnesses like heart disease takes up a significant portion of a family's total income and may be rather pricey. As a result, hospitalization is often necessary since the patients need extensive therapy for their condition.

Table 2. Estimation of total deaths attributable to cardiovascular disease between the years 2000 and 2030.

Country	**2020**		**2030**	
	Quantitative measure of morality.	Estimated annualized rate per 100,000.	Quantitative measure of mortality.	Estimated annualized rate per 100,000.
India	72,221,165	3172	16,937,070	2570
Brazil	104,840	2021	16,141,620	1857
China	5,656,890	1395	10,350,030	1653

1.2. Literature Review

Numerous studies have been conducted using machine learning, deep learning, fuzzy logic, and data mining tools and methodologies to predict cardiac disease. Researchers have employed a variety of datasets, algorithms, and procedures [7]; the findings they have seen thus far and future work will be used to determine the most effective techniques for diagnosing cardiovascular disease. The literature review has been divided into three categories on the basis of techniques including deep learning, machine learning, and ensemble learning.

1.2.1. Machine Learning Techniques

The major goal of this study is to create a predictive model with a small chance of success for people with heart disease. Weka was used for the experiment, and the Cleave Land UCI dataset underwent the following processing steps during training: pre-processing, classification, regression, clustering, association, and visualization. WEKA's Explorer mode is employed to try out classification tools. For the analysis, decision tree classifiers, including J48, the Logistic Model Tree Algorithm, and Random Forest, were used, along with 10-fold cross-validation and reduced error trimming. The j48 algorithm with decreased error pruning has the highest accuracy overall. If alternative discretization approaches, multiple classifiers, voting strategies, and additional decision tree algorithms (Gain ratio, Gini index) were used, the accuracy might be increased [8]. Taylan et al. proposed the prediction of heart disease using Naïve Bayes and k-means clustering. K-means clustering has been used to improve the efficiency of the desired output and is used for the grouping of attributes, followed by the Naïve Bayes algorithm for prediction of the disease. Naïve Bayes is mainly used when the inputs are generally high but it

gives no absurdity when the inputs are low. Moreover, it answered complex queries with respect to interpretation and accuracy [12]. Usama et al. [13] introduced a major purpose, to automate a method for diagnosing heart disease using historical data and information. The purpose was to discuss numerous knowledge abstraction methodologies leveraging data mining techniques, as well as their advantages and disadvantages. Data mining makes use of the Decision Tree algorithm, the Neural Network algorithm, and the Naïve Bayes algorithm. By calculating Shannon entropy, the ID3 Algorithm builds trees of decisions. Due to the constraints of building a short tree decision tree from set-off learning data, we use Quinlan's C4.5 and C5.O calculations to flow the data. Naïve Bayes surpasses the aforementioned strategies since it is not dependent on any input attributes. Future studies and the implementation of other ways to alleviate the problem of high overfitting are possible. In the investigation in [14], the subject of restricting and summarizing the many data mining approaches that are used in the field of medical prediction is investigated. The dataset was analyzed using the Naïve Bayes, KNN, and Decision Tree algorithms, which are all different supervised machine learning methods. The Tangra tool and 10-fold cross-validation are used in this step, and afterwards, the findings are compared in order to accomplish the categorization of the dataset. Thirteen different characteristics were employed for the comparison and contrast. The accuracy of the Decision Tree is the highest at 99.2%, followed by the accuracy of the Naïve Bayes method at 96.5% and the accuracy of the classification using clustering at 8.83% [14].

The major goal of another study [15] was to create a prediction model for the detection of heart disease using transthoracic echocardiography using data mining techniques. In the research, 7339 cases were extracted and pre-processed using data gathered at PGI, Chandigarh, between 2008 and 2011. Using the Weka 3.6.4 machine learning software, three unique supervised machine learning techniques were applied to a model of a pre-processed transthoracic echocardiography dataset. These methods are named J48 Classifier, Naïve Bayes, and Multilayer Perceptron. The distribution frequency was used to check for noise, inconsistency, and missing data, while box plots were used to locate outliers. The effectiveness of the models was evaluated using standard metrics, including accuracy, precision, recall, and F-measure. Random selection of training and test data samples was performed using ten cross-validations. With a classification accuracy of 95.56%, the J48 classifier is the most successful at predicting heart disease based on the specified parameters. To enhance classification accuracy and forecast particular kinds of heart disease, researchers will need to conduct additional tests with a larger number of datasets in the future [15]. The research entailed predicting cardiovascular disease using a different method. In order to forecast heart attacks successfully using data mining, the focus of this study is on the use of a wide variety of approaches, as well as combinations of a large number of target criteria.

A number of supervised machine learning strategies, such as Naïve Bayes and Neural Network, as well as a Weighted Association A Priori algorithm and a Decision algorithm, have been utilized for the purpose of conducting an analysis of the dataset. These strategies were used in order to accomplish this goal. For the purpose of another investigation [16], the data mining program Weka, version 3.6.6, was used. All of the tools that were necessary for the pre-processing of data, as well as classification, regression, clustering, association rules, and visualization, are linked to Weka. The Decision Tree algorithm has outperformed the others with an accuracy of 99.62% while using 15 criteria. In addition, the accuracy of the Decision Tree and Bayesian Classification algorithms increases even further when the genetic algorithm is used to minimize the actual data size in order to obtain the ideal subset of characteristics that are required for predicting heart disease. This is because the genetic algorithm is able to find the ideal subset of attributes that are necessary for making the prediction. This helps to ensure that the Decision Tree and Bayesian Classification algorithms produce the most accurate results possible. They used the Association Classification approach in addition to the prior algorithm and the Mafia algorithm [16].

Recently, Paladino et al. [17] proposed evaluating three AutoML tools (PyCaret, AutoGluon, and AutoKeras) on three datasets, comparing their performance with ten tradi-

tionally developed ML models. Traditionally designed models achieved 55–60% accuracy, while AutoML tools outperformed them. AutoGluon consistently achieved 78–86% accuracy, making it the top tool. PyCaret's performance varied (65–83%), and AutoKeras showed fluctuating results (54–83%). AutoML tools simplify ML model creation, potentially surpassing traditional methods.

1.2.2. Deep Learning Techniques

For the development of effective heart disease prediction [18], Multilayer Perceptron with a back propagation neural network is employed. The dataset contains 303 records, of which 40% is used for training the Multilayer Perceptron neural network, and the remainder is used for testing using the backpropagation algorithm. The experiment was conducted using Weka 3.6.11. Based on the actual desired output obtained from the Learning algorithm, in each neuron, the backpropagation algorithm calculates the error followed by the calculation of the output for each neuron, which in turn results in improved weights during the whole processing. The whole system developed shows an accuracy of 100% [18]. The reference [19] provided a novel idea to recognize individuals with a heart attack in color photos by identifying heart attack-related postures. Images of persons in infarction-free conditions and those who may be experiencing one were combined to create a dataset. Convolutional Neural Networks are used in the process of recognizing infarcts. A total of 1520 photos—760 of class "Infarct" and 760 of class "No Infarct"—were included in the first image data collection. These have been practised on a series of specifically created photos that feature people acting out heart attacks. The classification of infarcts shows encouraging results with 91.75% accuracy and 92.85% sensitivity.

Recently, Rani et al. [20] suggested a method, dubbed cardio help, that uses Convolutional Neural Networks, a type of deep learning algorithm, to predict the likelihood that a patient would have a cardiovascular illness (CNN). The suggested approach focused on modelling temporal data while applying CNN for early-stage HF prediction. They employed Convolutional Neural Networks and a cutting-edge dataset from the UCI library to predict a potential cardiac ailment (CNN). Some heart test parameters, as well as typical human behaviors, were included in the dataset. According to experimental findings, the proposed method performed better than the current methods in terms of performance evaluation measures. The proposed method's attained accuracy is 97%. The application of deep learning models was used to segment the left ventricle using magnetic resonance data. The results were very close to the actual situation. They suggested a different approach for automatically segmenting the left ventricle using magnetic resonance data and employed a deep learning strategy together with a level-set method. UNET was recommended by researchers [21] for segmenting biological images. To identify and differentiate boundaries in UNET, classification was performed on each individual pixel. In order to improve segmentation in medical imaging, the Fully Convolutional Network was modified and utilized as the basic architecture for the U-Net design. A technique incorporating a deep learning architecture has been proposed [22].

Another recent research [23] endeavour introduces a methodology utilizing a CNN model that effectively combines the strengths inherent in both dense and residual blocks. By capitalizing on the benefits of residual and dense connections, this model is able to enhance the flow of information, propagate gradients more effectively, and facilitate the reuse of features, ultimately resulting in improved performance. The proposed model was composed of interleaved residual-dense blocks, with the option of incorporating pooling layers for downsampling. Heartbeat classification into five distinct classes is accomplished by utilizing a linear support vector machine (LSVM), which simplifies the feature learning and representation derived from ECG signals. In order to address various issues such as baseline drift, power line interference, and motion noise, the initially collected ECG data are subjected to a denoising procedure. Subsequently, resampling techniques are employed to mitigate the impact of class imbalance. The proposed method underwent rigorous evaluation through extensive simulations conducted on well-established benchmarked

datasets, with a wide array of performance measures being utilized. On average, the proposed approach achieved an accuracy rate of 98.5%, a sensitivity rate of 97.6%, a specificity rate of 96.8%, and an AUC (Area Under the Curve) value of 0.99. In [24], an effective ML model was introduced for CKD prediction using pre-processing, feature selection, hyperparameter optimization, and ML techniques. On the UCI CKD dataset, the suggested model performed well with 100% accuracy. The study found that new pre-processing methods, the Boruta feature selection, the k-nearest neighbors technique, and grid-search cross-validation (CV) hyperparameter optimization improve early CKD identification. In another work, a machine learning model [25] was offered to predict cardiac disease using pre-processing, hyperparameter optimization, and ensemble learning methods. Using three comparable Kaggle datasets, a complete dataset for model evaluation was created. The strategy obtained 98.15% accuracy by using the additional tree classifier, normalizing the data, grid search cross-validation (CV) for hyperparameter optimization, and partitioning the dataset 80:20 for training and testing.

1.2.3. Ensemble Learning Techniques

Yang et al. [26] introduced a smote-XGboost-based approach to cardiovascular disease prediction in this study. A method was presented for feature selection that takes advantage of information gain; next, they used the hybrid Smote–Enn algorithm to handle imbalanced datasets. Lastly, the model was trained using the processed HDD dataset. XGboost was tested experimentally in comparison to five baseline algorithms. The results showed that the proposed model does very well across all four assessment metrics, with a prediction accuracy of 91.44%. Importantly, for the prediction of heart disease, they further quantify the feature relevance of the chosen algorithm [26]. To enhance the effectiveness of machine learning classifiers in predicting the risk of heart disease, the work used three attribute evaluator approaches to choose key features from the Cleveland heart dataset. By using the chi-squared attribute assessment approach, the SMO classifier accomplished an impressive feat. Over time, they found that, by carefully selecting attributes and fine-tuning the classifier's hyperparameters, the prediction performance was much enhanced. The study used a smaller dataset of 303 instances, three feature selection techniques, and ten machine learning classifiers. However, the classifier's performance is sufficient.

Numerous machine learning algorithms and feature selection approaches have a great deal of untapped potential. Reddy et al. [27] aimed to merge datasets in the future to increase the number of observations and run additional tests to enhance the classifier's predicting performance by choosing the right characteristics. The study created data mining algorithms with an 81.82% accuracy. They employed C4.5, CART, and RIPPER for the rule basis, compared the three fuzzy rule-based strategies in this research and deployed their system on 411 datasets [21]. The major goal of the research was to use a real-world dataset and several methods to categorize cardiac diseases. Predicting the existence of cardiac disease was carried out using the k-mode clustering technique on a dataset of patients. In order to prepare the dataset for analysis, the age attribute was converted to years and then divided into 5-year intervals. Similarly, the diastolic and systolic blood pressure values were divided into 10-interval bins. To account for the different features and development of heart disease in men and women, the dataset was additionally divided according to gender [28]. A recent study [29] compared some ensemble models, including XGBoost, CatBoost, Random Forest, and Extra Tree, in terms of classifying heart disease, and proposed that the best-performing model was Extra Tree with 97% accuracy.

1.3. Motivations and Challenges

The impetus behind conducting this research originates from the formidable global health challenge posed by cardiovascular diseases. As the foremost cause of mortality on a global scale, heart diseases necessitate innovative and precise predictive models for early identification and intervention. Acknowledging the limitations inherent in traditional

clinical methods, the study aims to navigate the intricacies and variability associated with cardiovascular conditions.

The primary objective is to construct a resilient and all-encompassing framework that leverages advanced machine learning techniques, explicitly focusing on boosting ensemble algorithms. The drive behind this lies in the potential of these techniques to heighten predictive accuracy, providing a more nuanced comprehension of the factors contributing to heart diseases. By exploring the synergies within boosting algorithms like AdaBoost, Catboost, and Gradient Boosting, the research strives to contribute to creating a sophisticated tool for early detection and risk assessment. Furthermore, the motivation extends to implementing preventive measures and timely interventions. In the face of an escalating global burden of heart disease, developing a practical predictive framework is pivotal in shaping public health strategies and enhancing patient outcomes. The research aspires to furnish healthcare professionals with a reliable tool for discerning individuals at risk, ultimately fostering proactive healthcare measures and mitigating the overall impact of cardiovascular diseases on global health. In summary, the motivation for this research is deeply rooted in the urgent necessity to augment our predictive capabilities for heart diseases, providing a proactive and productive approach to address this substantial public health concern.

2. Technical Details of Dataset

The heart disease dataset [30] offers an extensive range of attributes pertaining to cardiovascular health and lifestyle decisions, including specific patient information such as age, gender, cholesterol levels, blood pressure, heart rate, and factors such as diabetes, family history, smoking habits, obesity, and alcohol consumption. In addition, lifestyle variables such as the number of hours spent exercising, food habits, stress levels, and sedentary hours are taken into account. Medical factors, such as prior cardiac issues, drug use, and triglyceride levels, are taken into account. Factors such as income and geographical features, including nation, continent, and hemisphere, are also taken into account. The dataset contains 8763 patient records from various locations worldwide. It includes a vital binary classification feature that indicates the existence or absence of a heart attack risk. This dataset serves as a valuable resource for predictive analysis and research in the field of cardiovascular health. This dataset (See Table **??**) serves as a valuable resource for predictive analysis and research in the field of cardiovascular health.

Table 3. Description of heart disease dataset.

Index	Feature	Description	Value Type
1	Age	Age of the patient	Numerical
2	Sex	Gender of the patient	(Male/Female)
3	Cholesterol	Cholesterol levels of the patient	Numerical
4	Blood Pressure	Blood pressure of the patient (systolic/diastolic)	Numerical
5	Heart Rate	Heart rate of the patient	Numerical
6	Diabetes	Whether the patient has diabetes	(Yes/No)
7	Family History	Family history of heart-related problems	(1: Yes, 0: No)
8	Smoking	Smoking status of the patient	(1: Smoker, 0: Non-smoker)
9	Obesity	Obesity status of the patient	(1: Obese, 0: Not obese)
10	Alcohol Consumption	Level of alcohol consumption by the patient	(None/ Light/ Moderate/ Heavy)

Table 3. *Cont.*

Index	Feature	Description	Value Type
11	Exercise Hours Per Week	Number of exercise hours per week	Numerical
12	Diet	Dietary habits of the patient	(Healthy/ Average/ Unhealthy)
13	Previous Heart Problems	Previous heart problems of the patient	(1: Yes, 0: No)
14	Medication Use	Medication usage by the patient	(1: Yes, 0: No)
15	Stress Level	Stress level reported by the patient	(1–10)
16	Sedentary Hours Per Day	Hours of sedentary activity per day	Numerical
17	Income	Income level of the patient	Numerical
18	BMI	Body Mass Index (BMI) of the patient	Numerical
19	Triglycerides	Triglyceride levels of the patient	Numerical
20	Physical Activity Days Per Week	Days of physical activity per week	Numerical
21	Sleep Hours Per Day	Hours of sleep per day	Numerical
22	Country	Country of the patient	Numerical
23	Continent	Continent where the patient resides	Numerical
24	Hemisphere	Hemisphere where the patient resides	Numerical
25	Heart Attack Risk	Presence of heart attack risk	(1: Yes, 0: No)

Frequency Distribution

This experiment's dataset has a good mix of classes, with class 1 representing heart disease (4442 instances) and class 0 representing no heart disease (4321 instances), as shown in Figure 1. If the dataset is not balanced for the problem statement, machine learning and ensemble learning models will produce poor results. Sampling techniques can be used in certain situations to create a balanced dataset.

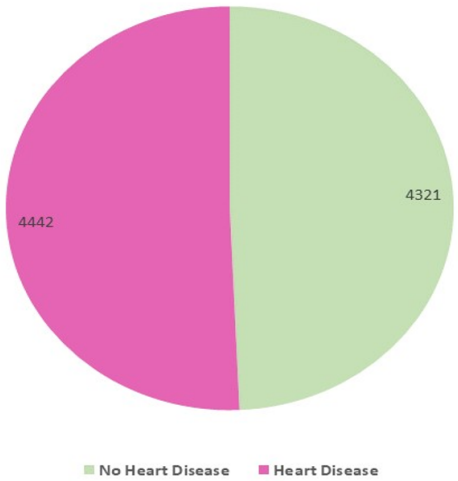

Figure 1. Frequency distribution of classes in the heart disease dataset.

The histogram is a significant statistical analysis technique for data visualization. It depicts the continuous variable distribution for a given interval of time. Through a histogram, the data are plotted by dividing them into sections called bins. The main use of the histogram is to inspect the underlying distribution of frequency, as in the case of normal distribution, skewness, outliers, and so on. Figure 2 below represents the histogram for each attribute separately to analyze their distribution.

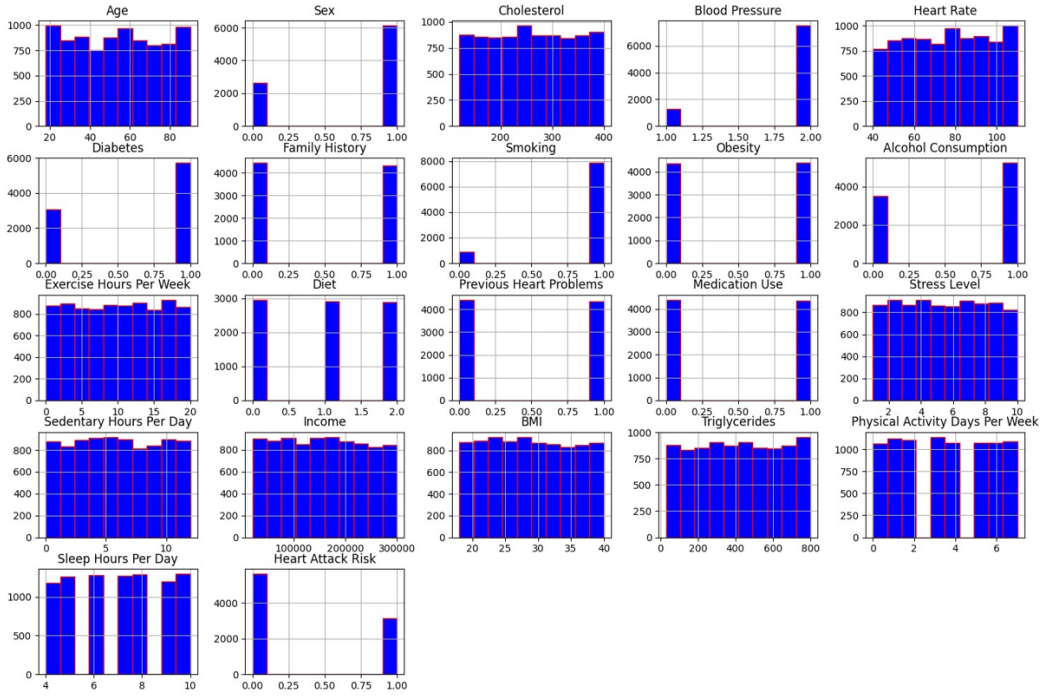

Figure 2. The distribution of the heart dataset features.

A map of the concentration of values between graphical figures, in which matrix values that are individual are presented as colors, is called a heat map. It is used in the visualization of two dimensions of the matrix shown in Figure 3. It is also helpful for pattern finding and hence gives a perspective of depth. In Figure 3, below, we used the color function in order to create colors of the heat map and added labels (row/column) to this. The heat map given below indicates how one attribute is co-related to another attribute, that is, whether it is negatively or positively correlated.

Descriptive statistics are crucial for defining the features of data and condensing the facts to facilitate human comprehension and interpretation. Table 4 provides an overview of the statistical measurements for the clinical characteristics, including the number of records, the lowest (min) value, the highest value, the maximum value, the mean, and the standard deviation (Std). As an example, the age property has a mean value of 53.84. The dataset has a mean value and a standard deviation of 21.21. The highest and lowest age values in the set are 90, and the ages are 18 years individually. These statistical metrics are likewise computed for the other 22 properties.

Table 4. Statistical analysis of the heart disease dataset.

Attribute	Count	Mean	Std	min	25%	50%	75%	max
Age	8763	58.8	21.2	18	35	54	72	90
Sex	8763	0.67	0.45	0	0	1	1	1
Cholesterol	8763	259	80	120	192	259	330	400
Blood Pressure	8763	1.85	0.35	1	2	2	2	2

Table 4. *Cont.*

Attribute	Count	Mean	Std	min	25%	50%	75%	max
Heart Rate	8763	75	20	40	58	75	93	110
Diabetes	8763	0.65	0.47	0	0	1	1	1
Family History	8763	0.49	0.49	0	0	0	1	1
Smoking	8763	0.896	0.304	0	1	1	1	1
Obesity	8763	0.501	0.5	0	0	1	1	1
Alcohol consumption	8763	0.589	0.49	0	0	1	1	1
Exercise hours/W	8763	10.01	5.78	0.002	4.98	10.06	15.05	19.99
Diet	8763	0.992	0.81	0	0	1	2	2
Previous heart problem	8763	0.49	0.5	0	0	0	1	1
Medication use	8763	0.498	0.5	0	0	0	1	1
Stress level	8763	5.46	2.85	1	3	5	8	10
Sedentary H/D	8763	5.99	3.46	0.001	2.998	5.93	9	11.9
Income	8763	158,263	80,575	20,062	88,310	157,866	227,749	299,954
BMI	8763	28.89	6.31	18	23.4	28.7	32.3	39.9
Triglycerides	8763	418.3	223.17	30	227	418	612	800
Physical act/w	8763	3.489	2.28	0	2	3	5	7
Sleep hours/day	8763	7.02	1.98	4	5	7	9	10
HA Risk	8763	0.358	0.47	0	0	1	1	1

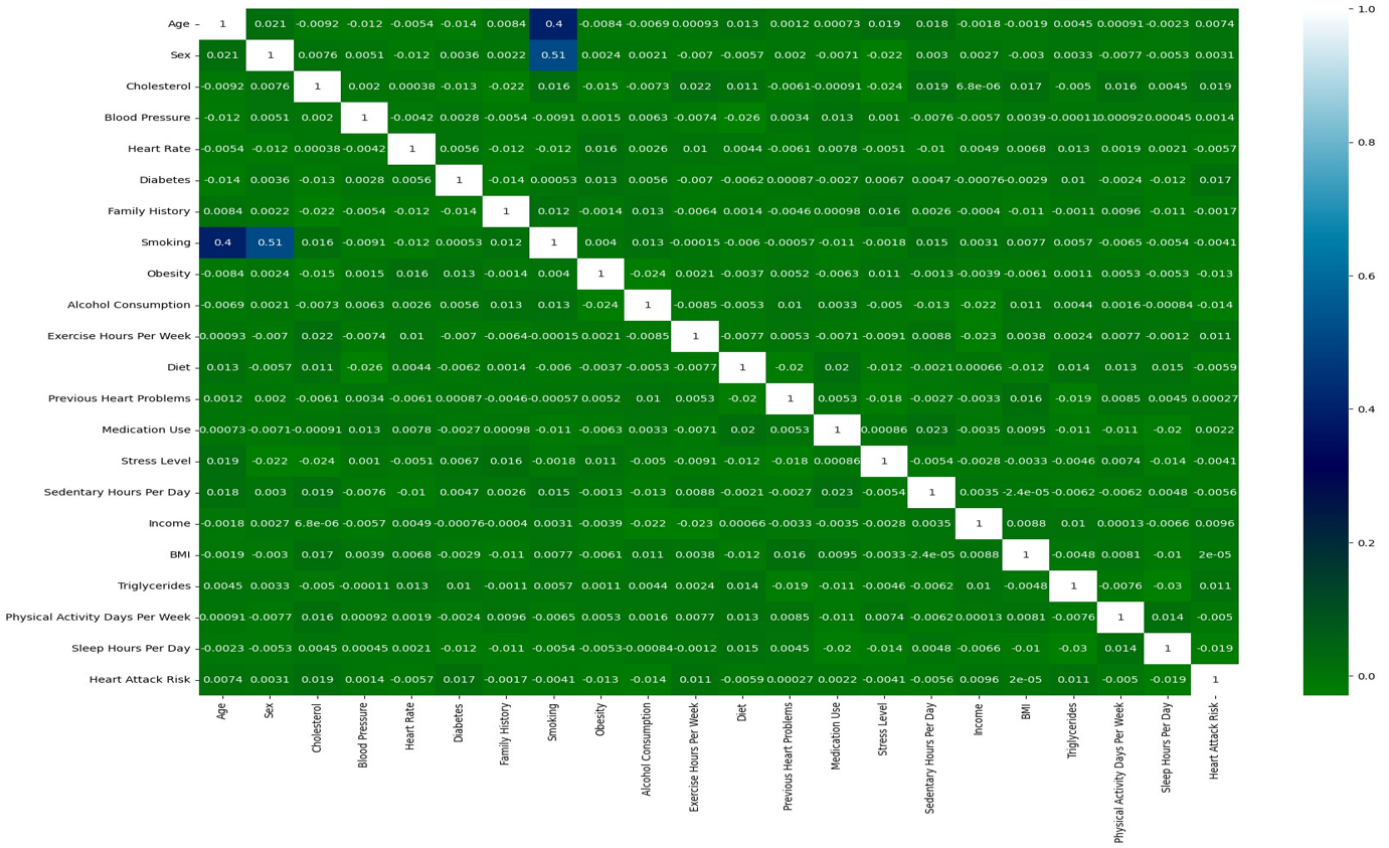

Figure 3. The relationships and patterns within the heart disease dataset.

3. Methods

The whole procedure, from data collection through the production of useful outcomes, is shown in Figure 4, and the technical details of the prediction model can be seen in Algorithm 1 below.

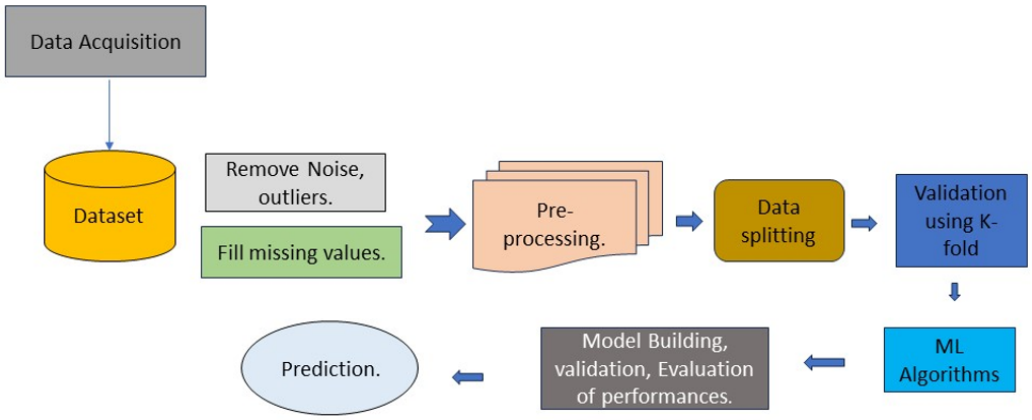

Figure 4. The details of methodology adapted.

Algorithm 1 Workflow of Methodology Employed:

- Input: Dataset
- Output: Prediction Model
- BEGIN
- Step 1; ingress dataset
- Step 2: pre-process the dataset (data transformation, data cleaning)
- Step 3: Xtrain, Ytrain...70% of dataset.
- Step 4: Xtest, Ytest.....30% of dataset.
- Step 5: D1 models and their Algorithms
- mn = (DT, RF, Boosting techniques)
- for $(i = 0, i < 3, i++)$ do
- Model.fit();
- Model.predict();
- Print (Accuracy(i), Confusion matrix, Roc curve)
- End
- Step 6: Placement of framework
- STOP

3.1. Machine Learning Framework

All experimental data were analyzed using Anaconda. Anaconda is a public and licensed accessible version of the popular Python and R language skills for scientific computing, designed to streamline deployment and package management (machine learning and data science applications, pre-processing of massive amounts of data, predictive analysis, etc.). Furthermore, we used TensorFlow to implement all ML models. TensorFlow is an open-source machine learning framework developed by Google. It provides tools for building and deploying machine learning models, including neural networks. TensorFlow supports distributed computing and offers high-level and low-level APIs. Keras is another Python library with a high-level neural network API built on top of TensorFlow. It simplifies the process of creating and training deep learning models with an easy-to-use syntax. Keras includes pre-trained models and supports various neural network architectures. Finally, scikit-learn, which is a popular ML library for Python, is applied. It offers a wide range of algorithms and tools for tasks like classification, regression, and clustering.

Scikit-learn follows a modular design and provides utilities for data pre-processing and model evaluation.

3.2. Pre-Processing

The pre-processing of data is an integral step in achieving improved outcomes prior to the development of ensemble machine learning models. Through the utilization of techniques such as resampling and discretization, the acquired dataset underwent pre-processing using the Integrated Development Environment Spyder with Python (3.9.1) as the programming tool.

The necessary libraries for data quality evaluation have been imported, and missing values have been filled. Although data imputation is not employed for every feature, we have filled the missing values pertaining to some attributes with their mean of neighboring values. Some attributes with the missing values were: age = 31, BP = 28, cholesterol = 23, BMI = 11, obesity = 05, and income = 08. The boxplot was used to find outliers, and the interquartile range approach was used to replace the outlier with viable sample values. However, out of 26 attributes, attributes like patient ID, country name, continent name, and hemisphere have been discarded. Before creating the machine learning models, data transformation was conducted to increase data efficiency. The interquartile range (IQR) encompasses the data points falling between the first and third quartiles, respectively. The outliers are removed through an interquartile range, setting a threshold of 1.5. Although it depends upon the type of dataset. Any points outside the range given are removed: Q1 − 1.5 × 1QR to Q3 + 1.5 × 1QR. The process of pre-processing is shown in Figure 5 below.

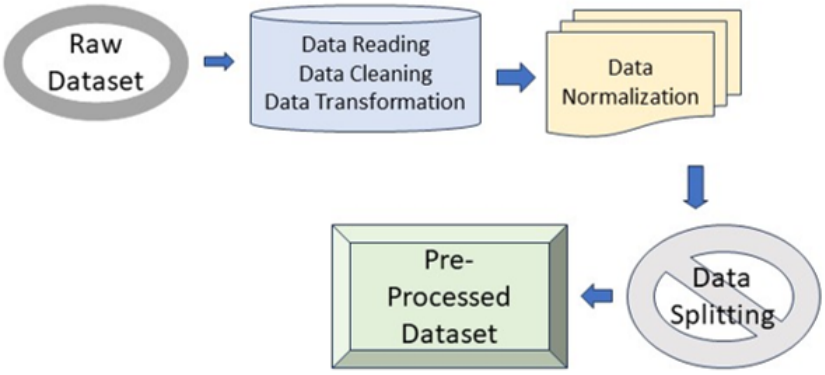

Figure 5. The details of pre-processing steps.

The dataset splitting ratio has been chosen with care, taking into account aspects such as overfitting, model complexity, size of the dataset, and the unique requirements of machine learning activity. The dataset is split into a 60:40 ratio. This means that, for every algorithm, 60% of the dataset is used for training, and the remaining 40% of the dataset is used for testing the algorithm. Through the train_test_split function, the implementation of the train–test–split evaluation procedure is carried out using the sci-kit learn library. The loaded dataset is taken by the function as the input, and then two subsets of this split dataset are returned. Ideally, we split the original dataset into input(x) and output(y) columns, followed by the function calling that passes both the arrays and splits them exactly into the train and test subsets separately. After the generation of results, 10-fold cross-validation will also be used to validate the results, and then all these results will be compared in terms of optimal accuracy.

3.3. Ensemble Learning Approaches

Ensemble learning approaches are widely investigated across several domains for the purpose of addressing practical challenges [31]. These models have achieved substantial advancements in the accurate forecasting, identification, assessment, and prediction of many medical conditions. This research examined three ensemble-learning-based boosting

algorithms for the purpose of predicting cardiac disease. The experiment made use of the algorithms listed below.

3.3.1. Decision Tree

When it comes to classification and regression, one non-parametric supervised learning approach that is often utilized is Decision Trees or DTs. The objective is to train a model to learn basic decision rules derived from data attributes in order to forecast the value of a target variable. A piecewise constant approximation is one way to look at a tree [32].

3.3.2. Random Forest

Random Forest algorithms consist of three primary hyperparameters that must be configured prior to the training process. These crucial parameters encompass the node size, the quantity of trees, and the number of features sampled. Once established, the Random Forest classifier becomes an invaluable tool for tackling both regression and classification problems. The underlying structure of the Random Forest algorithm is essentially a compilation of Decision Trees. Each individual tree within this ensemble is constructed by utilizing a data sample drawn from a training set with a replacement, which is commonly referred to as the bootstrap sample. To evaluate the effectiveness of this algorithm, one-third of the training sample is reserved as test data, known as the out-of-bag (oob) sample. This oob sample will play a significant role in our subsequent analysis and exploration of the Random Forest algorithm [33].

3.3.3. Gradient Boosting

Gradient boosting is a machine learning approach that belongs to the ensemble family. It is often used for both classification and regression applications. The algorithm constructs a prognostic model by assembling a collection of feeble learners, often Decision Trees, and then amalgamates them to generate a more robust and precise model. The fundamental concept behind Gradient boosting is to iteratively train further models in order to rectify the inaccuracies of the preceding models. The mathematical intricacies entail the use of calculus, specifically the calculation of gradients and partial derivatives. The choice of loss function will dictate the mathematical formulations for these gradients. Typical loss functions used in machine learning include mean squared error for regression problems and cross-entropy for classification tasks. The intricacies might become more intricate, but the underlying premise is iterative optimization to minimize the loss by modifying the model's predictions [34,35].

3.3.4. CatBoost

CatBoost [36] is a supervised machine learning technique used by the Train Using Auto ML program. It utilizes decision trees for both classification and regression tasks. CatBoost is characterized by two primary attributes: its ability to handle categorical data and its use of gradient boosting. Gradient boosting is an iterative approach that involves constructing many decision trees. Each successive tree enhances the outcome of the preceding tree, resulting in superior outcomes. CatBoost enhances the original gradient boost technique to achieve a more efficient implementation. CatBoost addresses a constraint present in other decision tree-based techniques, where it is usually necessary to preprocess the input by converting category text variables into numerical values, such as one-hot encodings. This technique has the capability to directly handle a mixture of category and non-categorical explanatory factors without the need for pre-processing. Preprocessing is an integral aspect of the algorithm. CatBoost employs an encoding technique known as ordered encoding to encapsulate category attributes [37].

3.3.5. XGBoost

It employs an ensemble of diverse decision trees (weak learners) to compute similarity scores independently. To address the issue of overfitting during the training phase,

the method adjusts the gradient descent and regularization procedure. You are able to control overfitting with the help of regularization, which is provided by XGBoost. This is accomplished by imposing L1/L2 penalties on the weights and biases of each tree [38].

After performing optimization techniques, the objective function of XGBoost is composed of two distinct components that serve to capture both the model's deviation and the regularization term, aiming to prevent overfitting. Let us denote the dataset as $D = \{(x_i, y_i)\}$, which encompasses a comprehensive collection of n samples and m features. Within this dataset, the predictive variable can be visualized as an additive model, comprising a combination of k fundamental models [39]. Upon conducting sample predictions, the outcome can be summarized as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \varphi \tag{1}$$

$$\varphi = \{f(x) = w_s(x)\}\left(s : R^m \to T, w_s \in R^T\right). \tag{2}$$

In the context where $\hat{y}_i$ symbolizes the label prediction and x_i stands for one of the samples, the predicted score is denoted by $f_k(x_i)$ or the given sample. Moreover, φ is used to signify the set of regression trees, which essentially encapsulates the tree structure parameters of s, $f(x)$. Additionally, w serves as a representation of the weight of the leaves and the number of leaves.

The objective function of XGBoost encompasses not only the traditional loss function but also the model complexity, thus enabling its utility in evaluating the operational efficiency of the algorithm. Within the framework of Formula (3), the first term is indicative of the traditional loss function, whereas the second term pertains to the model's complexity. Consequently, the comprehensive nature of the XGBoost objective function allows for a thorough assessment of both the predictive performance and the overall complexity of the model [39].

$$\text{Obj} = \sum_{i=1}^{m} l\left(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)\right) + \Omega(f_k), \tag{3}$$

$$\Omega(f_k) = \gamma T + 1/2\lambda w^2. \tag{4}$$

In the realm of these two exquisite formulas, the mystical entity denoted by the ethereal symbol i gracefully signifies the ethereal concept of the numerous samples that reside within the sacred dataset, while the enigmatic symbol m adorns itself with the mantle of representing the grandiose total amount of data that have been ceremoniously imported into the ethereal k th tree. As if imbued with arcane power, the enigmatic symbols γ and λ step forward to take on the audacious task of adjusting the very fabric of complexity within the ethereal tree. Behold, for in this realm, the very essence of regularization terms emerges as a mighty force, capable of adorning the final learning weight with a velvety smoothness and shielding it from the treacherous clutches of overfitting, like a knight in shining armor.

3.3.6. AdaBoost

As an ensemble method, AdaBoost [40] (or adaptive boosting) applies machine learning techniques. One-level decision trees, also known as decision trees with a single split, are the most used estimator in AdaBoost. Decision stumps are a common name for these trees. In this method, every piece of data is given the same weight while building a model. It then gives points that were wrongly classified as a higher weight. The following model gives more weight to all points that have higher weights. It will not stop training models until the reported error is reduced [25,41].

3.3.7. Light GBM

Distributed systems are ideal for maximizing Light GBM's performance [42]. With Light GBM, you may train decision trees that expand "leafwise", meaning that, for any given circumstance, you only divide the tree once based on the gain. Especially when working with smaller datasets, leaf-wise trees have the potential to overfit. Overfitting may be prevented by reducing the tree depth. Light GBM employs a histogram-based approach, which involves dividing data into bins based on the distribution's histogram. Instead of using each data point, the bins are used for iteration, gain calculation, and data splitting [43].

3.4. Performance Measures

The results yielded from the proposed algorithms may be evaluated by the following below-mentioned measures [24,44,45]:

Accuracy is interrupted from the given formula:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}. \tag{5}$$

In the equation, TP and TN stand for True Positive and True Negative, and FP and FN stand for False Positive and False Negative, respectively. TP + TN signifies the percentage of correctly classified instances, and TP + TN + FN + FP signifies the total number of correctly and incorrectly classified instances.

Precision is part of the significant instances involving the retrieved instances. The equation for precision is given below:

$$\text{Precision} = \frac{TP}{(TP + FP)}. \tag{6}$$

Recall is a small portion of appropriate instances, which is retrieved over the total quantity of relevant instances. The equation for the recall is given below:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{7}$$

The following equation, which satisfies the definition of specificity, is given below:

$$\text{Specificity} = \frac{TN}{(TN + FP)}. \tag{8}$$

F-measure is based on double the precision times recall divided by the sum of precision and recall. The equation for F-Measure is given below:

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

A confusion matrix, also referred to as an error matrix, is an incredibly useful tool in machine learning. It serves as a comprehensive and informative table that effectively summarizes the performance of a classification model. This matrix (See Figure 6) is extensively utilized to evaluate the accuracy and effectiveness of a model's predictions, providing valuable insights that aid in decision-making processes. Moreover, the confusion matrix acts as a visual representation that encompasses a wide range of information regarding how well a model classifies different classes. By comparing the predicted labels against the actual labels, it effectively showcases the model's capabilities and limitations. Structured in a square matrix format, the rows and columns correspond to the true and predicted classes, allowing for a detailed analysis of the model's performance. An essential aspect of the confusion matrix is its ability to facilitate the computation of various performance metrics. These metrics include accuracy, precision, recall, and F1 score, each offering unique insights into the model's predictive capabilities. By providing a clear distribution of correct and incorrect predictions across different classes, the confusion matrix offers a comprehensive

understanding of the model's strengths and weaknesses. Overall, the confusion matrix serves as an invaluable tool in the field of machine learning. Its ability to summarize and visualize the performance of a classification model provides researchers and practitioners with the necessary information to make informed decisions. One can significantly enhance the accuracy and effectiveness of their models by harnessing the power of the confusion matrix.

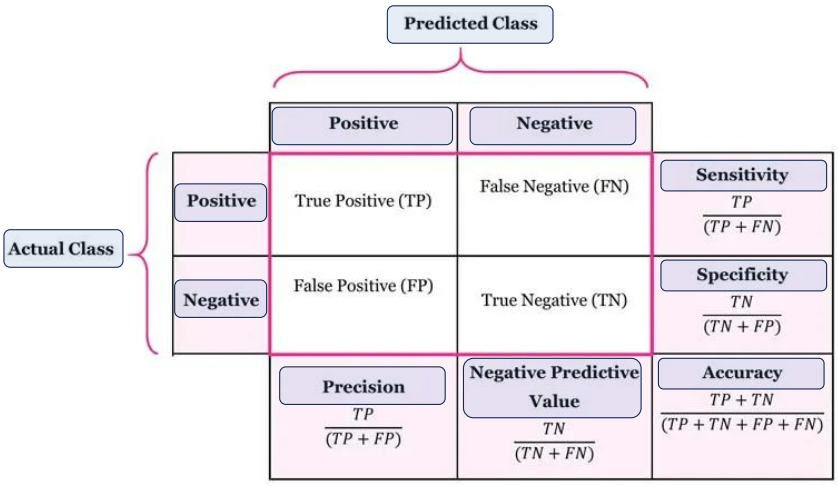

Figure 6. The components of the confusion matrix.

3.5. Validation

The next step is to randomize the supplied dataset and, for this, the k-fold cross-validation approach is utilized to test the data for the assessment of various ensemble machine learning algorithms. The dataset is partitioned into k subgroups containing the same data. This search employs 10-fold. As a result, data are 10-folded, with each fold around the same size. As a result, for each of the ten subsets of data, this validation approach employs one-fold for testing and the remaining nine-fold for training, as seen in Figure 7 below.

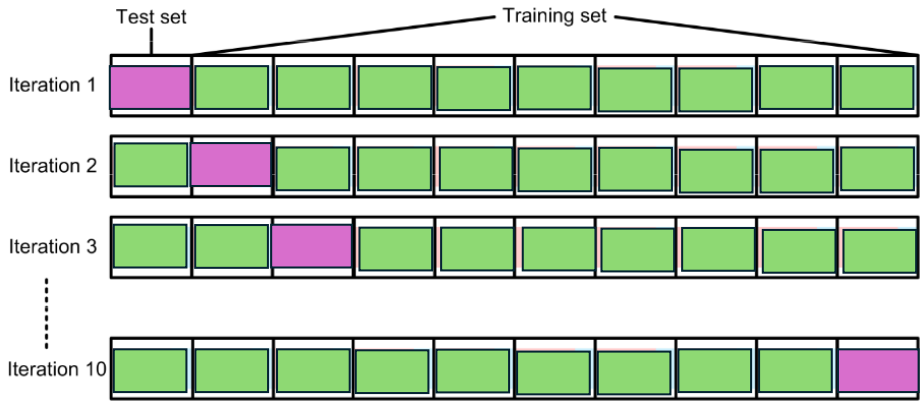

Figure 7. Ten-fold cross validation technique.

4. Experimental Results

The confusion matrix shown below is used to evaluate the performance of boosting models for identifying mislabeled/errors in predicting cardiac illness. It compares actual results to projected values using four factors: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The confusion matrix shown in Figure 8 below is used to evaluate the performance of boosting models for identifying mislabeled/errors in predicting heart illness. It compares actual results to projected values using four factors: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

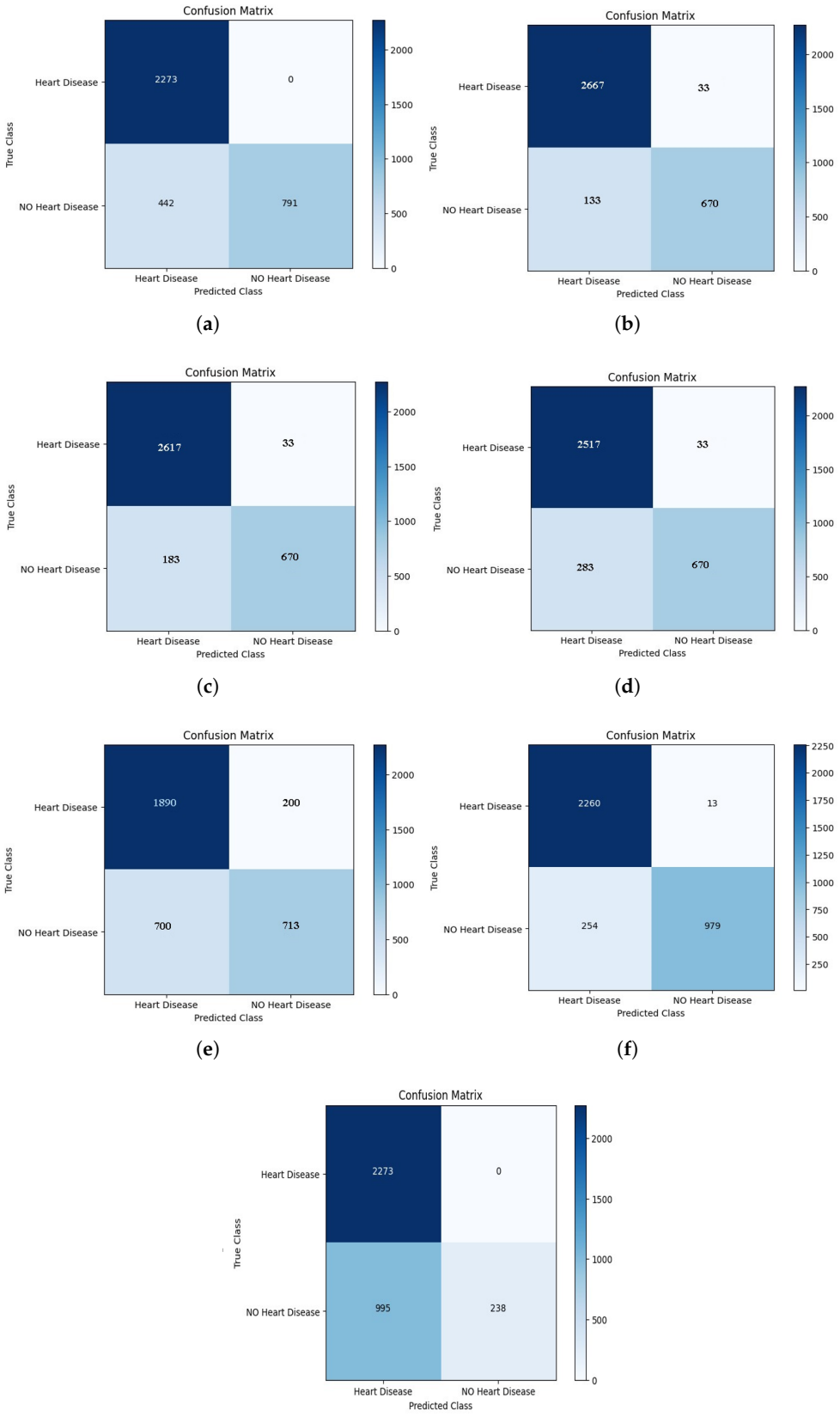

Figure 8. The testing confusion matrix of (**a**) GBoost, (**b**) AdaBoost, (**c**) CatBoost, (**d**) Light Boost, (**e**) Random Forest, (**f**) XGBoost, and (**g**) Decision Tree.

Testing Performance of Algorithms Employed: The accuracies of the algorithms are displayed in Table 5 below. Before pre-processing, the algorithms achieved a testing accuracy of 60.73%, 63%, 67%, 68%, 71.21%, 71.54%, and 72%. After pre-processing, the algorithms achieved an accuracy of 71%, 74%, 87%, 93%, 90%, 92.38%, and 95%, correspondingly.

Table 5. The average performance of seven ensemble learning models to classify heart disease.

Algorithm	Accuracy	Sensitivity	Specificity	Precision	NPV	FPR	FDR	FNR	FI-Score	MCC
DT	0.716	0.695	1	1	0.193	0	0	0.304	0.82	0.366
RF	0.743	0.729	0.78	0.904	0.504	0.219	0.095	0.27	0.807	0.457
GBoost	0.909	0.898	0.953	0.987	0.703	0.046	0.012	0.101	0.94	0.766
CatBoost	0.8739	0.8372	1	1	0.6415	0	0	0.1628	0.9114	0.7329
XGBoost	0.9238	0.899	0.9869	0.9943	0.794	0.0131	0.0057	0.101	0.9442	0.8356
Light GBM	0.938	0.934	0.953	0.987	0.785	0.046	0.012	0.065	0.96	0.828
AdaBoost	0.952	0.952	0.953	0.987	0.8344	0.0469	0.0122	0.0475	0.9698	0.8628

4.1. Feature Importance for AdaBoost

Feature importance is a technique used in machine learning to determine the most important features (i.e., variables or columns) in a dataset that contribute the most to the outcome of a model. The feature important scores are calculated for AdaBoost only because it attained the highest accuracy. The feature importance is calculated by setting a threshold of 0.3. If we are supposed to perform feature selection, we generally set a threshold and then only look at features with relevance ratings higher than that. Reducing overfitting and simplifying the model are two potential benefits of setting a threshold of 0.3. However, it varies from model to model. The feature importance for XGB is shown in Figure 9 below: Also, so far as the feature selection technique is concerned, we have used a correlation-based feature selection technique. The results have been validated through five-fold and ten-fold cross-validation.

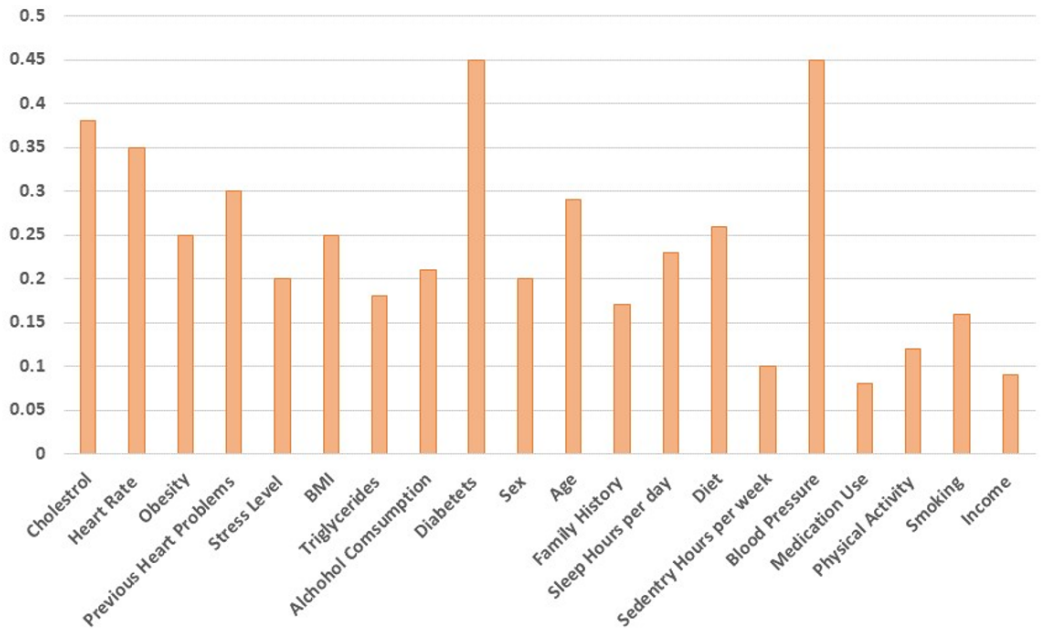

Figure 9. Features importance based on the AdaBoost.

4.2. Comparative Results

Table 6, presented below, serves as a means of conducting an analytical comparison between our proposed framework and the existing body of work. This comparison is rooted in the examination of various crucial aspects, including the methodology employed, the dataset utilized, and the analysis conducted. It is worth noting that the majority of lifestyle factors that were studied and explored in the context of our proposed framework are shared by all the studies that were conducted for the purpose of comparison. Through meticulous analysis, it was revealed that our proposed framework had exhibited a remarkable performance, particularly in terms of various assessment criteria, with a notable focus on the accuracy of heart disease prediction. In order to surpass the achievements of previous relevant studies and to ensure the attainment of superior outcomes, a series of advanced procedures were implemented. These procedures encompassed techniques such as data imputation, which was employed to manage missing values effectively, as well as the identification and replacement of outliers through the utilization of the well-regarded boxplot method. Moreover, in order to enhance the reliability and validity of the findings, the data underwent a rigorous process of standardization and normalization, which was carried out using the transformation method. Furthermore, to further solidify the credibility of the results produced by our proposed framework, the K-fold cross-validation approach was meticulously employed during the development and implementation stages of the suggested framework. This approach ensures that the findings obtained are both robust and generalizable, thereby enhancing the overall validity of the framework's predictions and conclusions.

Table 6. The average performance of seven ensemble learning models to classify heart disease.

Ref	Year	Method	Dataset	Splitting Ratio	Results
[46]	2020	CNN	UCI with 303 instances and 14 attributes	70:30	Accuracy = 94.78%
[47]	2022	AB, LR, CART, SVM, LDA, RF, XGB	UCI with 303 instances and 14 attributes	70:30	XGB Precision = 90, Recall = 100, F-measure = 95, and Accuracy = 91.80%.
[8]	2023	SVR, ANFIS, M5 Tree	UCI with 1028 instances and 13 attributes	70:30	ANFIS, ANN-LM, and ANN-BFG, Accuracy = 94.7%, 96.2%, and 91.50.
[48]	2022	Stacking ensemble of LR, RF, SGD, Ensemble of GDC and ADA	Cleveland, Switzerland, long beach Va Satlog	60:40	MTE, Accuracy, Precision, Recall, F-measure, AUC-ROC = 91.84%, 91.75%, 95.22%, 93.30% and 94.05%.
[49]	2023	RF, LR, NB, DT	UCI	70:30	Accuracy, ROC, Precision, Recall, F1, and Score = 0.98%, 0.63%, 0.98%, 1.00%, 0.99%.
[19]	2021	CNN	UCI	60:40	Precision = 0.8669, Recall = 0.8174, Fl-score = 0.8414, Accuracy = 0.8667.
Our Model		DT, RF, CatB, GB, XGB, Adaboost, Light GBM	Kaggle with 8763 and 26 attributes	60:40	AdaBoost, Accuracy, Sensitivity, Specificity, and Precision = 0.952%, 0.952%, 0.953%, and 0.987%.

5. Conclusions

This paper makes a significant contribution by comparing ensemble machine learning classifiers, mainly focusing on boosting algorithms, for the early prediction of cardiovascular disease (CVD). Pre-processing steps were implemented to enhance dataset quality, emphasizing the management of corrupted and missing information, as well as outlier elimination. In addition to the AdaBoost model, six other ensemble learning algorithms were employed to predict CVD, and their performances were evaluated using a range of statistical indicators. The experimental results highlight the AdaBoost model's exceptional accuracy, with a 96% accuracy in the training set and 95% in the testing set. The robustness

of the DT, RF, CatBoost, GBoost, XGBoost, AdaBoost, and Light GM algorithms was validated through five-fold and ten-fold cross-validation techniques. AdaBoost consistently demonstrated a superior performance across various criteria, including accuracy, recall, and f1-score. Furthermore, the study incorporated a correlation feature-based selection technique. To enhance the predictive potential of boosting classifiers for heart disease, future efforts should involve training and testing on larger, primary datasets. The study acknowledges the need for further research, from data collection to result in visualization, to elevate the research's calibre. In the future, extending this study to develop a web application based on boosting and other machine learning, as well as deep learning models, using larger datasets could yield optimal results for efficient heart disease prediction. Additionally, exploring other deep learning techniques, particularly those incorporating image data, could further advance the understanding and prediction capabilities in the realm of cardiovascular health. This research sets the stage for future investigations, emphasizing the continual evolution and expansion of methodologies for enhanced heart disease prediction and management.

Author Contributions: The authors' contributions are as follows. Conceptualization, N.N., S.J. and M.N.; methodology, N.N., S.J.; software, N.N. and S.J.; validation, N.N. and S.J.; formal analysis, N.N. and S.J.; investigation, N.N., S.J. and M.N.; resources, N.N.; data curation, N.N. and S.J.; writing—original draft preparation, N.N. and S.J.; writing—review and editing, N.N., S.J. and M.N.; visualization, N.N.; supervision, S.J.; project administration, S.J.; funding acquisition, M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The heart disease dataset used in this study can be found at https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset (10 November 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Petrelli, A.; Gnavi, R.; Marinacci, C.; Costa, G. Socioeconomic inequalities in coronary heart disease in Italy: A multilevel population-based study. *Soc. Sci. Med.* **2006**, *63*, 446–456. [CrossRef]
2. Sharma, H.; Rizvi, M. Prediction of heart disease using machine learning algorithms: A survey. *Int. J. Recent Innov. Trends Comput. Commun.* **2017**, *5*, 99–104.
3. Gheorghe, A.; Griffiths, U.; Murphy, A.; Legido-Quigley, H.; Lamptey, P.; Perel, P. The economic burden of cardiovascular disease and hypertension in low-and middle-income countries: A systematic review. *BMC Public Health* **2018**, *18*, 975. [CrossRef] [PubMed]
4. Ahmad, G.N.; Ullah, S.; Algethami, A.; Fatima, H.; Akhter, S.M.H. Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection. *IEEE Access* **2022**, *10*, 23808–23828. [CrossRef]
5. Mohammad, F.; Al-Ahmadi, S. WT-CNN: A Hybrid Machine Learning Model for Heart Disease Prediction. *Mathematics* **2023**, *11*, 4681. [CrossRef]
6. Osisanwo, F.Y.; Akinsola, J.E.T.; Awodele, O.; Hinmikaiye, J.O.; Olakanmi, O.; Akinjobi, J. Supervised machine learning algorithms: Classification and comparison. *Int. J. Comput. Trends Technol. (IJCTT)* **2017**, *48*, 128–138.
7. Rashid, Y.; Bhat, J.I. Topological to deep learning era for identifying influencers in online social networks: A systematic review. *Multimed. Tools Appl.* **2023**, 1–44. [CrossRef]
8. Taylan, O.; Alkabaa, A.S.; Alqabbaa, H.S.; Pamukçu, E.; Leiva, V. Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods. *Biology* **2023**, *12*, 117. [CrossRef]
9. Adeli, A.; Neshat, M. A fuzzy expert system for heart disease diagnosis. In Proceedings of the International Multi-Conference of Engineers and Computer Scientists, IMECS 2010, Hong Kong, 17–19 March 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1–6.
10. Neshat, M.; Zadeh, A.E. Hopfield neural network and fuzzy Hopfield neural network for diagnosis of liver disorders. In Proceedings of the 2010 5th IEEE International Conference Intelligent Systems, London, UK, 7–9 July 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 162–167.
11. Neshat, M.; Yaghobi, M.; Naghibi, M.B.; Zadeh, A.E. Fuzzy expert system design for diagnosis of liver disorders. In Proceedings of the International Symposium on Knowledge Acquisition and Modeling, Wuhan, China, 21–22 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 252–256.

12. Li, X.; Zhao, Y.; Zhang, D.; Kuang, L.; Huang, H.; Chen, W.; Fu, X.; Wu, Y.; Li, T.; Zhang, J.; et al. Development of an interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: Findings of the US NHANES from 2003 to 2018. *Chemosphere* **2023**, *311*, 137039. [CrossRef]
13. Usama, M.; Qadir, J.; Raza, A.; Arif, H.; Yau, K.L.A.; Elkhatib, Y.; Hussain, A.; Al-Fuqaha, A. Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE Access* **2019**, *7*, 65579–65615. [CrossRef]
14. Ngiam, K.Y.; Khor, W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **2019**, *20*, e262–e273. [CrossRef] [PubMed]
15. Nissa, N.; Jamwal, S.; Mohammad, S. Early detection of cardiovascular disease using machine learning techniques an experimental study. *Int. J. Recent Technol. Eng.* **2020**, *9*, 635–641. [CrossRef]
16. Kecman, V. Support vector machines—An introduction. In *Support Vector Machines: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 1–47.
17. Paladino, L.M.; Hughes, A.; Perera, A.; Topsakal, O.; Akinci, T.C. Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction. *AI* **2023**, *4*, 1036–1058. [CrossRef]
18. Rojas-Albarracin, G.; Chaves, M.Á.; Fernandez-Caballero, A.; Lopez, M.T. Heart attack detection in color images using convolutional neural networks. *Appl. Sci.* **2019**, *9*, 5065. [CrossRef]
19. Mehmood, A.; Iqbal, M.; Mehmood, Z.; Irtaza, A.; Nawaz, M.; Nazir, T.; Masood, M. Prediction of heart disease using deep convolutional neural networks. *Arab. J. Sci. Eng.* **2021**, *46*, 3409–3422. [CrossRef]
20. Rani, M.; Bakshi, A.; Gupta, A. Prediction of Heart Disease Using Naïve bayes and Image Processing. In Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Shimla, India, 26–28 November 2021; IEEE: Piscataway, NJ, USA, 2020; pp. 215–219.
21. Rairikar, A.; Kulkarni, V.; Sabale, V.; Kale, H.; Lamgunde, A. Heart disease prediction using data mining techniques. In Proceedings of the 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, 23–24 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8.
22. Zakariah, M.; AlShalfan, K. Cardiovascular Disease Detection Using MRI Data with Deep Learning Approach. *Int. J. Comp. Electr. Eng.* **2020**, *12*, 72–82. [CrossRef]
23. Ahmed, A.E.; Abbas, Q.; Daadaa, Y.; Qureshi, I.; Perumal, G.; Ibrahim, M.E. A Residual-Dense-Based Convolutional Neural Network Architecture for Recognition of Cardiac Health Based on ECG Signals. *Sensors* **2023**, *23*, 7204. [CrossRef]
24. Arif, M.S.; Mukheimer, A.; Asif, D. Enhancing the early detection of chronic kidney disease: A robust machine learning model. *Big Data Cogn. Comput.* **2023**, *7*, 144. [CrossRef]
25. Chandrasekhar, N.; Peddakrishna, S. Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes* **2023**, *11*, 1210. [CrossRef]
26. Yang, J.; Guan, J. A heart disease prediction model based on feature optimization and smote-Xgboost algorithm. *Information* **2022**, *13*, 475. [CrossRef]
27. Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Appl. Sci.* **2021**, *11*, 8352. [CrossRef]
28. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective heart disease prediction using hybrid machine learning technique. *South Asian J. Eng. Technol.* **2022**, *12*, 123–130. [CrossRef]
29. Asif, D.; Bibi, M.; Arif, M.S.; Mukheimer, A. Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. *Algorithms* **2023**, *16*, 308. [CrossRef]
30. Banerjee, S. Heart Attack Risk Prediction Dataset. 2023. Available online: https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset (accessed on 10 January 2024).
31. Hassan, C.A.u.; Iqbal, J.; Irfan, R.; Hussain, S.; Algarni, A.D.; Bukhari, S.S.H.; Alturki, N.; Ullah, S.S. Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors* **2022**, *22*, 7227. [CrossRef] [PubMed]
32. Tayefi, M.; Tajfard, M.; Saffar, S.; Hanachi, P.; Amirabadizadeh, A.R.; Esmaeily, H.; Taghipour, A.; Ferns, G.A.; Moohebati, M.; Ghayour-Mobarhan, M. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. *Comput. Methods Programs Biomed.* **2017**, *141*, 105–109. [CrossRef]
33. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* **2019**, *7*, 81542–81554. [CrossRef]
34. Kubat, M.; Kubat, J. *An Introduction to Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 2.
35. Graczyk, M.; Lasota, T.; Trawiński, B.; Trawiński, K. Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. In Proceedings of the Intelligent Information and Database Systems: Second International Conference, ACIIDS, Hue City, Vietnam, 24–26 March 2010; Part II 2; Springer: Berlin/Heidelberg, Germany, 2010; pp. 340–350.
36. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–12.
37. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for big data: An interdisciplinary review. *J. Big Data* **2020**, *7*, 94. [CrossRef]
38. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA; 13–17 August 2016; pp. 785–794.
39. Qiu, Y.; Zhou, J.; Khandelwal, M.; Yang, H.; Yang, P.; Li, C. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Eng. Comput.* **2021**, *38*, 4145–4162. [CrossRef]

40. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; Volume 96, pp. 148–156.
41. Ganie, S.M.; Dutta Pramanik, P.K.; Mallik, S.; Zhao, Z. Chronic kidney disease prediction using boosting techniques based on clinical parameters. *PLoS ONE* **2023**, *18*, e0295234. [CrossRef] [PubMed]
42. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–12.
43. Ceylan, Z.; Bulkan, S.; Elevli, S. Prediction of medical waste generation using SVR, GM (1, 1) and ARIMA models: A case study for megacity Istanbul. *J. Environ. Health Sci. Eng.* **2020**, *18*, 687–697. [CrossRef] [PubMed]
44. Chang, V.; Bhavani, V.R.; Xu, A.Q.; Hossain, M. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthc. Anal.* **2022**, *2*, 100016. [CrossRef]
45. Neshat, M.; Ahmedb, M.; Askarid, H.; Thilakaratnee, M.; Mirjalilia, S. Hybrid Inception Architecture with Residual Connection: Fine-tuned Inception-ResNet Deep Learning Model for Lung Inflammation Diagnosis from Chest Radiographs. *arXiv* **2023**, arXiv:2310.02591.
46. Sajja, T.K.; Kalluri, H.K. A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network. *Rev. D'Intelligence Artif.* **2020**, *34*, 601–606. [CrossRef]
47. Saboor, A.; Usman, M.; Ali, S.; Samad, A.; Abrar, M.F.; Ullah, N. A method for improving prediction of human heart disease using machine learning algorithms. *Mob. Inf. Syst.* **2022**, *2022*, 1410169. [CrossRef]
48. Hera, S.Y.; Amjad, M.; Saba, M.K. Improving heart disease prediction using multi-tier ensemble model. *Netw. Model. Anal. Health Informatics Bioinform.* **2022**, *11*, 41. [CrossRef]
49. Pandey, S. The Cardiovascular Disease Prediction Using Machine Learning. *Buana Inf. Technol. Comput. Sci. (BIT CS)* **2023**, *4*, 24–27. [CrossRef]