algorithms

MDPI

Article

Heart Disease Prediction Using Concatenated Hybrid Ensemble Classifiers

Annwesha Banerjee Majumder [1], Somsubhra Gupta [2], Dharmpal Singh [3], Biswaranjan Acharya [4], Vassilis C. Gerogiannis [5], Andreas Kanavos [6,*] and Panagiotis Pintelas [7,*]

1 Department of Information Technology, JIS College of Engineering, Kalyani 741235, India; annwesha.banerjee@jiscollege.ac.in
2 Department of Computer Science and Engineering, Swami Vivekananda University, Kolkata 700121, India; director.rnd@svu.ac.in
3 Department of Computer Science and Engineering, JIS College of Engineering, Kalyani 741235, India; hod_cse@jisuniversity.ac.in
4 Department of Computer Engineering AI, Marwadi University, Rajkot 360003, India; biswaranjan.acharya@marwadieducation.edu.in
5 Department of Digital Systems, University of Thessaly, 41500 Larissa, Greece; vgerogian@uth.gr
6 Department of Informatics, Ionian University, 49100 Corfu, Greece
7 Department of Mathematics, University of Patras, 26500 Patras, Greece
* Correspondence: akanavos@ionio.gr (A.K.); ppintelas@gmail.com (P.P.)

Abstract: Heart disease is a leading global cause of mortality, demanding early detection for effective and timely medical intervention. In this study, we propose a machine learning-based model for early heart disease prediction. This model is trained on a dataset from the UC Irvine Machine Learning Repository (UCI) and employs the Extra Trees Classifier for performing feature selection. To ensure robust model training, we standardize this dataset using the StandardScaler method for data standardization, thus preserving the distribution shape and mitigating the impact of outliers. For the classification task, we introduce a novel approach, which is the concatenated hybrid ensemble voting classification. This method combines two hybrid ensemble classifiers, each one utilizing a distinct subset of base classifiers from a set that includes Support Vector Machine, Decision Tree, K-Nearest Neighbor, Logistic Regression, Adaboost and Naive Bayes. By leveraging the concatenated ensemble classifiers, the proposed model shows some promising performance results; in particular, it achieves an accuracy of 86.89%. The obtained results highlight the efficacy of combining the strengths of multiple base classifiers in the problem of early heart disease prediction, thus aiding and enabling timely medical intervention.

Keywords: Adaboost; decision tree; extra trees classifier; hybrid ensemble voting classifier; K-nearest neighbor; logistic regression; machine learning; Naive Bayes

Citation: Majumder, A.B.; Gupta, S.; Singh, D.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A.; Pintelas, P. Heart Disease Prediction Using Concatenated Hybrid Ensemble Classifiers. *Algorithms* **2023**, *16*, 538. https://doi.org/10.3390/a16120538

Academic Editor: Frank Werner

Received: 19 October 2023
Revised: 15 November 2023
Accepted: 21 November 2023
Published: 25 November 2023

1. Introduction

Heart disease is a formidable global menace that claims a substantial number of lives annually. Early detection of this critical illness plays a pivotal role in preventing fatalities. However, timely and effective medical intervention for heart disease often encounters significant obstacles, which are attributable to restricted availability of medical assistance, a situation that is not confined solely to less economically developed countries [1]. These challenges also extend to undeserved communities within economically developed regions, emphasizing the widespread and multifaceted nature of constrained access to proper healthcare services. In response, Machine Learning (ML) models have emerged as potent tools to tackle these challenges, capitalizing on their capacity to analyze extensive datasets and make insightful predictions for many illness conditions, including heart disease [2,3]. By harnessing the power of ML algorithms, the relevant models hold the potential to

predict heart disease at its incipient stages, facilitating timely medical intervention and potentially saving lives.

The main motivation behind the current research study is deeply embedded in the imperative need to combat the overwhelming global prevalence of heart disease. This disease remains one of the leading causes of mortality worldwide, affecting individuals across various age groups, socioeconomic backgrounds, and geographical locations. Its pervasive impact extends beyond specific demographics and regions, underscoring the urgent need to develop robust and reliable early detection methods. By addressing the complexities associated with the timely identification and management of heart disease, this research seeks to contribute significantly to the global efforts to reduce the associated morbidity and mortality rates. By emphasizing the universal nature of this health challenge and its multifaceted repercussions, this study aims to underscore the critical need for comprehensive and effective predictive models that can transcend demographic and geographic boundaries. According to the World Health Organization (WHO), cardiovascular diseases are the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually. These statistics underscore the pressing need for effective heart disease prediction models that can aid in early diagnosis and intervention [4].

In particular, in this paper, we propose a machine learning-based model for the early prediction of heart disease. Our objective is to accurately assess the likelihood of heart disease based on patient data, with the ultimate aim of enhancing early detection, prognosis, and facilitating effective medical intervention and patient care. The proposed model is trained on a dataset sourced from the UC Irvine ML Repository (UCI) [5]. For the critical task of feature selection, we employ the Extra Trees Classifier. This classifier boasts several advantages, including robustness to noisy data, low bias and variance, computational efficiency, feature importance estimation, ability to handle interactions and nonlinear relationships, and resistance to overfitting [6]. These attributes make Extra Trees Classifier a valuable asset in the realm of feature selection in ML. To ensure the reliability of the model training, we standardize the dataset using StandardScaler. This preprocessing technique offers a multitude of benefits, including preserving the distribution shape, mean removal, variance scaling, mitigation of outlier impact, facilitation of feature comparison, improved convergence of optimization algorithms, and compatibility with various ML algorithms. Such advantages make StandardScaler a suitable tool for data standardization in ML pipelines, contributing to more robust and effective model training and evaluation [7].

For classification, we introduce an innovative approach: a concatenated hybrid ensemble voting classifier. This technique involves the fusion of two hybrid ensemble classifiers, each leveraging the advantages of a variety of base classifiers, including Support Vector Machine, Decision Tree, K-Nearest Neighbor, Logistic Regression, Adaboost, and Naive Bayes. By embracing ensemble classifiers, the proposed model achieves enhanced accuracy, robustness against noise and outliers, improved interpretability, and the ability to handle intricate relationships within the data [8]. The concatenation of these two ensemble classifiers offers several advantages, including superior performance, enhanced generalization capabilities, robustness against noise and outliers, improved handling of complex relationships, increased interpretability, adaptability to diverse datasets, and opportunities for customization and optimization.

In particular, our contributions entail a comprehensive refinement of existing heart disease prediction models through a systematic approach. First, we present a refined feature selection methodology, carefully integrating the Extra Trees Classifier to enhance the predictive accuracy of the selected features. Second, the present research study incorporates an advanced data standardization technique, leveraging the power of StandardScaler to ensure optimal data preprocessing. Third, we introduce a concatenated hybrid ensemble voting classifier, which is uniquely designed to leverage the strengths of ensemble classifiers while minimizing their inherent limitations. Finally, our research study delves into an extensive exploration of the benefits associated with classifier concatenation, thereby enhancing the interpretability and overall performance of the predictive model. Collectively,

these contributions establish an advancement in the domain of heart disease prediction by using ML algorithms, fostering improved accuracy, heightened robustness, and a deeper understanding of the interpretability of the ML methods.

The rest of the paper is organised as follows: Section 2 provides an overview of recent research in disease prediction using ML, serving as motivation for the current research work. Section 3 delves into the details of our model, encompassing the dataset, feature selection via the Extra Trees Classifier, data standardization using StandardScaler, and the concatenated hybrid ensemble voting classifier. Section 4 showcases the experimental outcomes, demonstrating the model's performance in predicting heart disease. Finally, Section 5 concludes the paper, summarizing the research findings and elucidating their implications.

2. Related Work

Several ML models have been proposed for heart disease prediction. The authors of [9,10] employed ensemble methods, incorporating classifiers like Naive Bayes, Decision Tree, Logistic Regression, and Random Forest to enhance prediction accuracy. Also, a hybrid model combining Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, and Decision Tree classifiers to improve accuracy and robustness was introduced in [11]. In a recent study [12], five different ML models were compared using a standard cardiovascular disease dataset. Among these models, bootstrap aggregation yielded the highest accuracy of 97.67%. Another noteworthy work focused on explainable AI [13] for cardiovascular disease prediction; the authors employed various traditional ML algorithms to enhance the models' interpretability.

Triguero et al. [14] introduced a comprehensive taxonomy, providing a detailed categorization based on the fundamental attributes exhibited in these approaches. Additionally, their extensive investigation delved into the classification performance across various datasets, elucidating the efficacy of these methodologies in a rigorous manner. In [15], the researchers introduce a novel semi-supervised learning algorithm founded on the principles of self-training. This algorithm employs multiple independent base learners initially and dynamically identifies the most promising base learner during the training phase using a strategy based on the number of highly confident predictions from unlabeled data.

The research studies in [16,17] explored the integration of various classifiers into their heart disease prediction models, including Random Forest, Support Vector Machine, Naive Bayes, and Decision Trees. In both studies, the Support Vector Machine (SVM) demonstrated the best performance in terms of accuracy and predictive capabilities. Furthermore, refs. [18,19] focused on feature selection and evaluation. Specifically, ref. [18] used Random Forest in combination with Logistic Regression to assess the importance of features, while [19] employed multiple classifiers, including K-Nearest Neighbors, Decision Tree, Logistic Regression, Naive Bayes, SVM, and Random Forest, to improve accuracy and performance.

A comprehensive analysis utilizing Logistic Regression, LightGBM, XGBoost, Gaussian Naive Bayes, SVM, and Gaussian Naive Bayes, achieving varying accuracies, was presented in [20]. Another ensemble model combining logistic regression and a majority voting approach, which outperformed existing methods for heart disease prediction, was proposed in [21]. In another work [22], Random Forest, Neural Network, Decision Tree, and SVM were combined in a hybrid model for heart disease prediction. The authors of [23] focused on using Deep Neural Network and S2 statistics to predict heart disease while addressing overfitting and underfitting issues. The authors of [24] achieved higher accuracy by working on an Artificial Neural Network for cardiovascular disease prediction. Decision Tree, Random Forest, and Naive Bayes classifiers were used in [25] for disease prediction. A hybrid approach combining Random Forest and Linear Model for heart disease prediction was utilized in [26].

In addition, an estimation model using bagging techniques with base learners including Naive Bayes, K-Nearest Neighbors, and Logistic Regression, was proposed in [27]. A detailed analysis of different ML algorithms for predicting cardiovascular disease, including

Random Forest, Decision Tree, Logistic Regression, SVM, and K-Nearest Neighbors, was conducted in [28]. A model that utilized SVM and Artificial Neural Network (ANN) was also presented in [29]. Moreover, an intelligent model utilizing Random Forest for disease prediction, focusing on the importance of smart disease prediction within a ML framework, was introduced in [30]. Finally, authors the authors of [31] proposed a model using Logistic Regression, Decision Tree, SVM, and Naive Bayes for heart disease prediction.

In [32], the researchers assessed the effectiveness of an ensemble semi-supervised learning technique in categorizing chest X-rays related to tuberculosis. Through multiple experiments, which are validated by statistical nonparametric tests, this study showcases the algorithm's competence, suggesting that dependable and resilient predictive models can be constructed even when utilizing a limited number of labeled data points alongside a substantial amount of unlabeled data. Similarly, the authors of [33] assess the effectiveness of two ensemble semi-supervised learning methods in the domain of credit scoring. The numerical experiments conducted reveal that the suggested algorithms surpass their individual semi-supervised learning counterparts, indicating that the integration of ensemble techniques within the semi-supervised learning framework can lead to the creation of dependable and resilient prediction models.

While studies like those described above have made significant contributions to the field of applying ML to heart disease prediction, they also exhibit certain limitations. Many of them rely on a limited set of classifiers and may not fully exploit the potential of ensemble learning. Moreover, feature selection techniques in some studies lack thoroughness, potentially leading to sub-optimal predictive performance. To address these limitations, the current research work introduces an approach that combines a wide range of base classifiers and employs advanced feature selection techniques, ultimately enhancing the accuracy and robustness of the proposed heart disease prediction model.

3. Proposed Model

The proposed model has been developed by amalgamating multiple hybrid ensemble models. The block diagram of this approach is depicted in Figure 1 below.

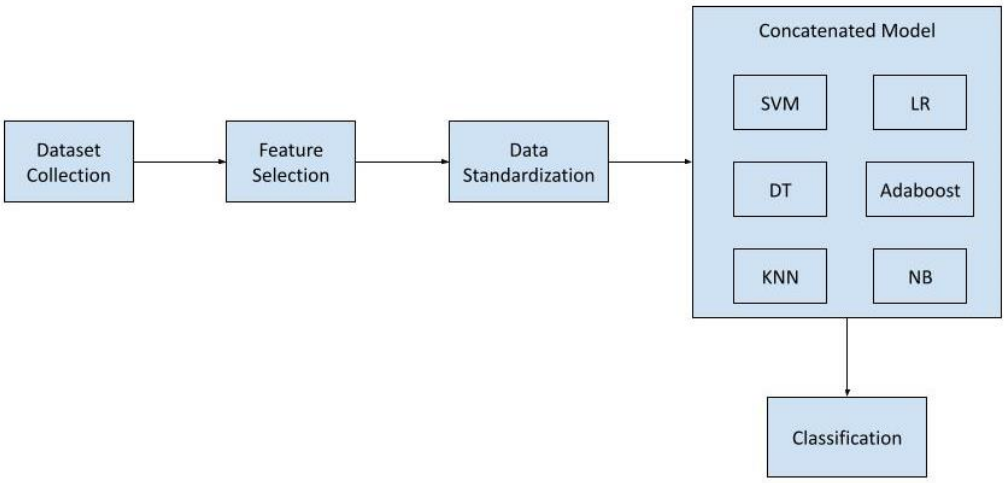


Figure 1. Block Diagram of Proposed Model.

To provide a detailed understanding of the classification process within the concatenated hybrid ensemble voting classifier, a process flowchart has been incorporated. This flowchart, presented in Figure 2, complements the block diagram in Figure 1 by illustrating the sequential steps involved in the decision-making process.

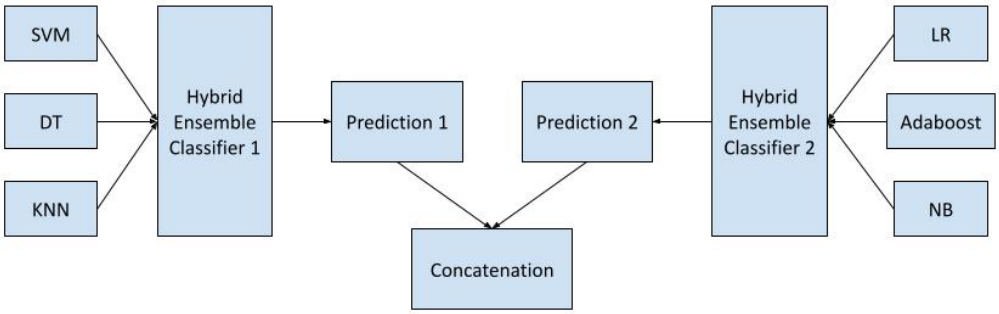


Figure 2. Flowchart of concatenated hybrid ensemble voting classifier.

3.1. Dataset

The dataset used in this study was obtained from the UCI Heart Disease Data Repository [34]. It comprises a total of 14 features, with the target variable being the dependent variable. The independent variables include age, sex, cp (chest pain), trestbps (resting blood pressure), chol (serum cholesterol), fbs (fasting blood sugar), restecg (resting electrocardiographic results), thalach (maximum heart rate achieved), exang (exercise-induced angina, with zero representing absence and one representing presence), oldpeak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment), ca (number of major vessels colored by fluoroscopy), and thal (thalassemia).

Before constructing the model, a comprehensive analysis and visualization of the dataset were conducted to gain valuable insights into the distribution of values. This initial exploratory analysis facilitated a deeper understanding of the data, allowing us to make informed decisions during the subsequent modeling phase. It is important to note that "angina" (exercise-induced angina) represents whether or not the patient experienced angina during the stress test, with values of zero indicating no angina and one indicating the presence of angina.

Table 1 presents the distribution of different chest pain values in relation to the target variable.

Table 1. Sample Chest Pain Distribution against Target.

Chest Pain Distribution	Percentage
Asymptomatic with Heart Disease	45.6%
Asymptomatic without Heart Disease	17.1%
Angina with Heart Disease	3.07%
Non-Angina without Heart Disease	30.3%
Atypical Angina with Heart Disease	3.05%

Table 2 displays the sample distribution of high and low blood sugar values against the target variable.

Table 2. Sample blood sugar distribution against target.

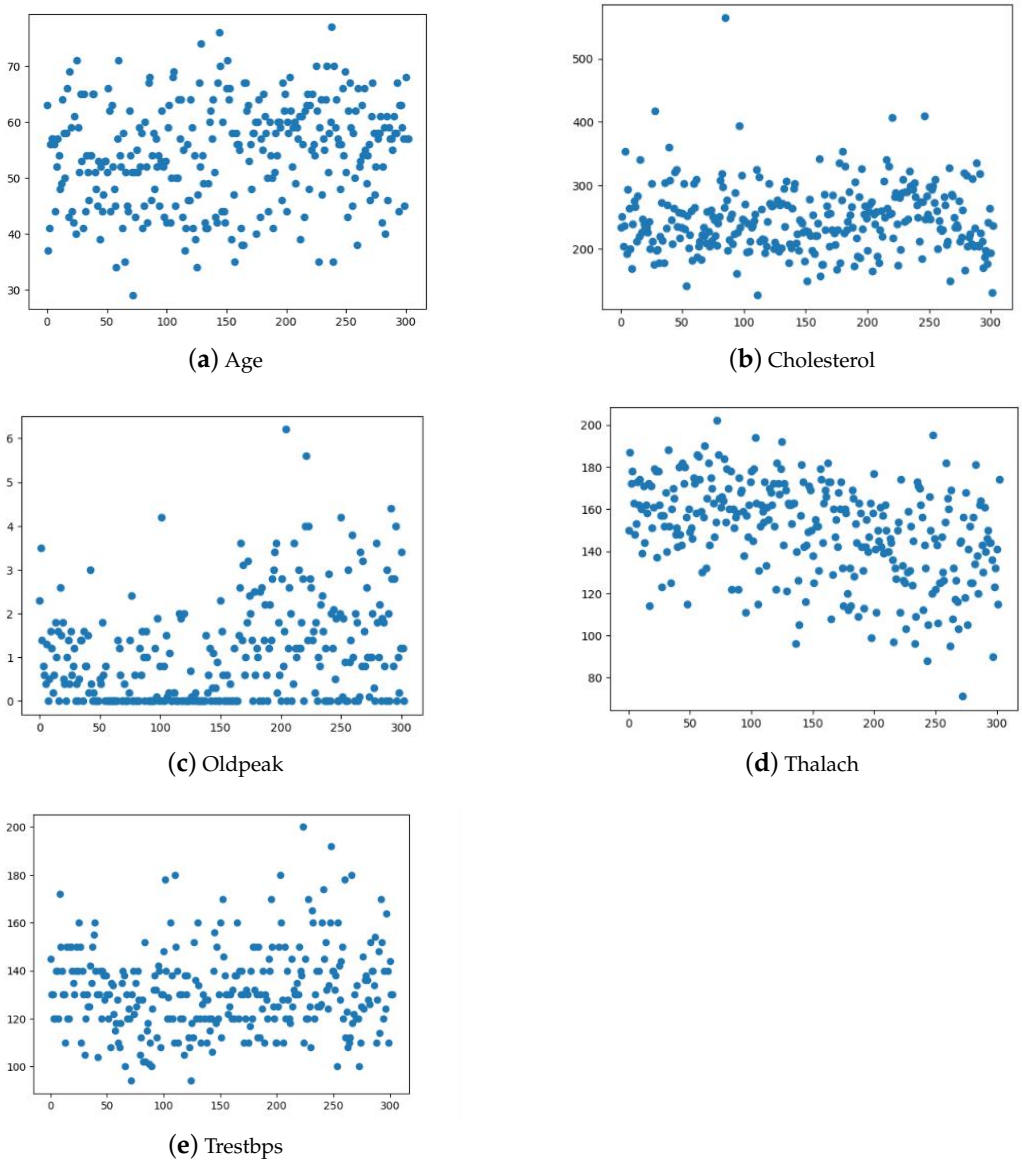
Blood Sugar Distribution	Percentage
High Blood Sugar with Heart Disease	48.89%
High Blood Sugar without Heart Disease	51.11%
Low Blood Sugar with Heart Disease	44.96%
Low Blood Sugar without Heart Disease	55.04%

In Table 3, the sample distribution of exang (exercise-induced angina) values with respect to the target variable can be depicted.

Table 3. Sample exang distribution against target.

Blood Sugar Distribution	Percentage
Feeling Angina during stress Test with Heart Disease	76.77%
Feeling Angina during stress Test without Heart Disease	23.23%
Not Feeling Angina during stress Test with Heart Disease	30.39%
Not Feeling Angina during stress Test without Heart Disease	69.61%

The distribution of continuous variables, including age, cholesterol, oldpeak, thalach, and trestbps, is visualized through the scatter plots in Figure 3. This figure helps us identify potential data patterns and outliers, which can influence the choice of appropriate modeling techniques and preprocessing steps.



(**a**) Age (**b**) Cholesterol (**c**) Oldpeak (**d**) Thalach (**e**) Trestbps

Figure 3. Distribution of age, cholesterol, oldpeak, thalach and trestbps.

3.2. Data Cleaning

Before training the ML model, it is essential to handle missing data, if applicable. Fortunately, this dataset was clean, with no missing values. However, in real-world scenarios, data cleaning often includes addressing missing data, outliers, and inconsistencies.

3.3. Feature Engineering

Feature engineering is a crucial step in preparing the data for ML. In the used dataset, some features were categorical, such as 'chest pain type' and 'thalassemia', while others were numerical, like 'age' and 'resting blood pressure'. To make the data suitable for applying ML models, we applied one-hot encoding to categorical features. Additionally, we employed the Extra Trees Classifier for feature selection. This method ranks the importance of each feature, allowing us to focus on the most informative attributes [35]. Feature selection not only reduces dimensionality, but also enhances model interpretability and generalization.

3.4. Data Standardization

Standardization is a critical preprocessing step, especially when working with algorithms that are sensitive to feature scaling, such as Support Vector Machines and K-Nearest Neighbors. We used the StandardScaler from scikit-learn v.1.2.2 to standardize the numerical features, transforming them to have a mean of zero and a standard deviation of one. This process ensures that all features have the same scale, preventing some features from dominating the others during model training:

The standard score z is defined as:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the mean of the training samples (or zero if with_mean = False), and σ is the standard deviation of the training samples (or one if with_std = False). This formula represents the process of transforming the input data x into standard scores, allowing for a mean of zero and a standard deviation of one.

3.5. Feature Selection

Feature selection plays a critical role in ML models, as not all features may positively contribute to decision-making. To address this, the Extra Tree classifier has been employed to select the most relevant features from the dataset. The Extra Trees Classifier, also known as Extremely Randomized Trees, is an ensemble learning method based on decision trees. Mathematically, we can represent the decision function of the Extra Trees Classifier as follows in Equation (2):

$$f(x) = C_k(x) \tag{2}$$

where $f(x)$ represents the decision function and $C_k(x)$ denotes the class assigned to the input feature vector x by the k-th decision tree.

The strategic feature selection process using the Extra Trees Classifier enhances the efficiency, interpretability, and generalization of our heart disease prediction model. Before feature selection, the dataset comprised a comprehensive set of features, including age, sex, chest pain type (cp), resting blood pressure (tresbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved during exercise (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), and thalassemia type (thal), as depicted in Table 4.

Table 4. Features before and after feature selection.

Name of Features before Feature Selection	Name of Features after Feature Selection
Independent Features: age, sex, cp, tresbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	Independent Features: age, cp, thalach, exang, oldpeak, slope, ca, thal
Dependent Feature: Target	Dependent Feature: Target

After careful consideration, the feature selection process retained the most informative features, including age, chest pain type (cp), maximum heart rate achieved during exercise (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), and thalassemia type (thal). This selection aligns with existing medical knowledge about factors that influence heart disease, ensuring that our model focuses on the most relevant aspects to achieve accurate predictions.

The selected features contribute significantly to the model's predictive capabilities while eliminating redundancy and reducing the risk of overfitting. This strategic feature selection process not only improves the computational efficiency of our model, but also enhances its interpretability and generalization to new data. Certain features, such as sex, thalach, and exang, were deemed less clinically relevant to heart disease prediction and were discarded during the feature selection process.

To visualize the significance of each feature, we apply the Extra Tree classifier, as shown in Figure 4. Notably, slope, age, oldpeak, thalach, thal, exang, ca, and cp were identified as the most influential features in our model.

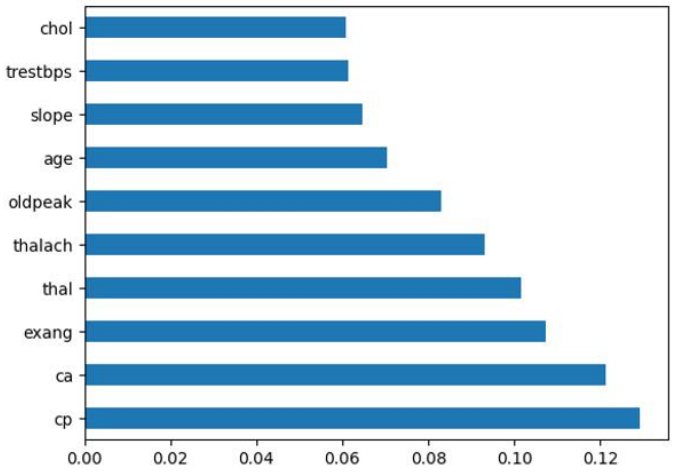


Figure 4. Feature importance applying extra tree classifier.

These insights into feature selection aim to provide a deeper understanding of the variables influencing our model's predictions.

3.6. Classifier Selection

The proposed heart disease prediction model leverages ensemble learning, combining multiple base classifiers to improve predictive accuracy and robustness. In our study, we select a diverse set of base classifiers, including:

- Support Vector Machine (SVM): SVM is an effective ML technique that is commonly employed for problems involving classification and regression [36,37]. The algorithm operates by identifying the hyperplane that achieves the highest degree of separation between distinct classes of data, while simultaneously increasing the margin between these classes. SVM exhibits notable efficacy in high-dimensional areas and possesses the ability to handle both linear and non-linear associations among data points. The technology in question is adaptable and extensively employed across diverse disciplines such as image recognition, text classification, and bioinformatics. Consequently, it holds significant value as an essential asset for data scientists and practitioners of ML.
- Decision Tree: The Decision Tree method is a fundamental ML technique that facilitates decision-making processes by organizing data into a hierarchical structure that resembles a tree [38]. The method is extensively employed for tasks involving classification and regression, offering a clear and comprehensible approach to gen-

erating predictions based on input data. Decision trees are frequently employed in sophisticated ensemble techniques, rendering them a valuable asset in the field of ML.

- K-Nearest Neighbor (KNN): KNN is a versatile classification and regression technique [39]. The 'k' nearest data points in the training set are used to produce predictions in KNN. In regression, the average neighbor value predicts the query point's class label, while in classification, the majority class determines it. KNN works in many applications, since it does not make any assumptions about data distributions.
- Logistic Regression: Logistic Regression is a widely employed statistical and ML model utilized for binary classification assignments [40,41]. Contrary to its nomenclature, this method is not employed for regression purposes; rather, it is utilized to estimate the likelihood of an input being classified into one of two distinct categories. This process involves the utilization of a sigmoid function to model the data, resulting in a mapping of input properties to a probability score ranging from 0 to 1. Logistic Regression is characterized by its simplicity, interpretability, and computational efficiency, rendering it a valuable analytical technique across many fields such as medical diagnosis, spam detection, and credit scoring. The linear model possesses the capability to be expanded to accommodate multi-class categorization, rendering it highly adaptable in practical applications.
- Adaboost: AdaBoost, also known as Adaptive Boosting, is a ML ensemble method employed to enhance the efficacy of weak classifiers [42]. The algorithm operates by iteratively training a sequence of weak classifiers, assigning higher importance to the cases that were incorrectly identified by the preceding classifiers. This approach directs the succeeding classifiers towards the more difficult cases, resulting in a robust and precise model. AdaBoost's adaptability allows it to handle various types of healthcare data, including clinical, genetic, or imaging data. Its feature selection capability can help prioritize important factors in disease diagnosis, and its reduced risk of overfitting ensures robust performance even in the presence of noisy or incomplete data. Overall, AdaBoost is a valuable tool for improving the accuracy and reliability of disease prediction models, which can have a significant impact on early diagnosis and treatment planning.
- Naive Bayes: The Naive Bayes algorithm is a widely utilized probabilistic classification technique that is known for its simplicity and effectiveness in the fields of ML and natural language processing [43]. The Naive Bayes classifier is derived from Bayes' theorem and relies on the assumption of conditional independence of characteristics. Despite the adoption of this simplifying assumption, Naive Bayes consistently demonstrates an impressive performance in various tasks such as text categorization and spam detection. The calculation of the probability of a data point's membership in a specific class renders it a handy instrument for decision-making and categorization purposes. The approach demonstrates computational efficiency and is particularly well-suited for the analysis of high-dimensional data, which may pose challenges for alternative algorithms.

By incorporating a variety of classifiers, we aim to capture different decision boundaries and patterns in the data, making the proposed model more resilient to noise and variability.

3.7. Ensemble Learning

The proposed approach introduces an ensemble learning technique, which combines the predictions of two hybrid ensemble classifiers:

1. Hybrid Ensemble 1: This ensemble consists of SVM, Decision Tree, and KNN classifiers. Each base classifier is trained on the preprocessed dataset independently.
2. Hybrid Ensemble 2: This ensemble includes Logistic Regression, Adaboost, and Naive Bayes classifiers, each trained on the same dataset as in Hybrid Ensemble 1.

The selection of base learners was conducted based on their robust performance in previous studies and their relevance to the specific characteristics of the dataset. While the ensemble method employs a majority voting scheme, the fusion of diverse classifiers with distinct decision boundaries enables the exploration of complementary aspects of the data, thereby enhancing the model's predictive capabilities. The inclusion of these basic classifiers in components 1 and 2 was chosen to leverage their individual strengths and ensure a diverse range of learning strategies within the hybrid ensemble framework.

Specifically, in designing Hybrid Ensemble 1, we aimed to integrate classifiers with diverse capabilities to enhance the model's overall performance. Three base classifiers were selected based on their individual strengths:

1. SVM: Linear kernels were chosen for their simplicity and robustness to linearly separable data. SVMs are known for producing efficient decision boundaries.
2. Decision Tree: Selected for its ability to represent nonlinear relationships in the data and for its interpretability.
3. KNN: Employed for recognizing local patterns and adjusting to the structure of the data.

The integration of these three classifiers in Hybrid Ensemble 1 provides the model with the ability to handle linear and nonlinear patterns, contributing to its generalization and robustness.

For Hybrid Ensemble 2, three distinct base classifiers were chosen:

1. Logistic Regression: A straightforward yet powerful linear classifier suitable for binary classification tasks.
2. AdaBoost: An ensemble technique known for building a powerful classifier by combining weak ones, adapting to complex data.
3. Naive Bayes: A probabilistic classifier frequently used in various domains, particularly in text categorization.

This ensemble design ensures a combination of classifiers with diverse attributes, enhancing the model's adaptability to different data features and improving the overall prediction accuracy. The rationale behind the selection aligns with established practices in the literature, promoting transparency and reproducibility of the proposed approach.

The predictions from both hybrid ensembles are then concatenated and used as input to a Voting Classifier. This final step aggregates the predictions from all base classifiers, employing a majority voting scheme to make the final prediction. The main reason behind this concatenated hybrid ensemble approach is to exploit the diversity of base classifiers, each with its strengths and weaknesses. By combining two hybrid ensembles, we aim to enhance the model's overall predictive performance.

The proposed hybrid ensemble classifier can be represented using the equation below (Equation (3)):

$$f(x) = mode(C_1(x), C_2(x), \ldots, C_N(x)) \tag{3}$$

where $f(x)$ represents the decision function of the ensemble voting classifier, $mode()$ returns the most frequent class among the predictions, and $C_i(x)$ represents the class predicted by the i-th base classifier for the input feature vector x.

For the first ensemble classifier, $C_1(x)$ represents the Support Vector Machine, $C_2(x)$ is the Decision Tree classifier, and $C_3(x)$ is the K-Nearest Neighbor classifier.

The decision function of the Support Vector Machine ($C_1(x)$) can be expressed as follows:

$$C_1(x) = sign(\sum \alpha_i y_i K(x_i, x) + b) \tag{4}$$

where $C_1(x)$ is the decision function, $sign()$ is the sign function returning $+1$ for positive values and -1 for negative values, $\sum$ represents the summation over all support vectors, α_i are the Lagrange multipliers (coefficients obtained during training), y_i is the class label of the i-th support vector ($+1$ or -1), $K(x_i, x)$ is the kernel function calculating the similarity between the i-th support vector x_i and the input feature vector x, and b is the bias term.

The Decision Tree classifier ($C_2(x)$) for the first ensemble classifier can be expressed as follows:

$$C_2(x) = c_1 \text{ if } T_1(x), c_2 \text{ if } T_2(x), \ \ldots, \ c_n \text{ if } T_n(x) \tag{5}$$

where $C_2(x)$ is the decision function of the decision tree, $c_1, c_2, \ldots, c_n$ are the class labels associated with the terminal nodes (leaves) of the decision tree, and $T_1(x), T_2(x), \ldots, T_n(x)$ are the decision conditions or rules based on the input feature vector x that guide the traversal of the decision tree.

The K-Nearest Neighbor classifier ($C_3(x)$) for the first ensemble classifier can be represented as follows:

$$C_3(x) = mode(C_1(x), C_2(x), \ldots, C_k(x)) \tag{6}$$

where $C_3(x)$ is the decision function of the KNN classifier, $mode()$ returns the most frequent class among the k nearest neighbors, and $C_i(x)$ represents the class label of the i-th nearest neighbor to the new data point x.

For the second ensemble classifier, $C_1(x)$ represents Logistic Regression, $C_2(x)$ is the Adaboost classifier, and $C_3(x)$ is the Naive Bayes classifier.

The decision function of Logistic Regression ($C_1(x)$) can be described as:

$$C_1(x) = \sigma(w \cdot x + b) \tag{7}$$

where $C_1(x)$ is the decision function of the Logistic Regression classifier, $\sigma()$ is the logistic function (sigmoid function), w is the weight vector, x is the input feature vector, $\cdot$ denotes the dot product, and b is the bias term.

The decision function of the Adaboost classifier ($C_2(x)$) can be expressed as:

$$C_2(x) = \sum a_i h_i(x) \tag{8}$$

where $C_2(x)$ is the decision function of the Adaboost classifier, $\sum$ represents the summation over all weak classifiers, a_i are the weights assigned to each weak classifier, and $h_i(x)$ represents the prediction of the i-th weak classifier for the input feature vector x.

The decision function of the Naive Bayes classifier ($f(x)$) can be defined as:

$$f(x) = \text{argmax}(c) P(c) \prod P(x_i|c) \tag{9}$$

where $f(x)$ is the decision function of the Naive Bayes classifier, argmax(c) returns the class c that maximizes the expression, $P(c)$ is the prior probability of class c, $P(x_i|c)$ is the conditional probability of feature x_i given class c, and $\prod$ represents the product operator, which calculates the product of the conditional probabilities for all features.

In the final step, these two ensemble classifiers are merged to create the ensemble classifier:

$$final(x) = g(C_1(x), C_2(x), \ldots, C_N(x)) \tag{10}$$

where $final(x)$ is the decision function of the concatenated classifier, $C_1(x), C_2(x), \ldots, C_n(x)$ represent the individual predictions of the base classifiers for the input feature vector x, and $g()$ is the final classifier that takes the concatenated feature vector as input and makes the final prediction.

In the proposed model, $C_1(x)$ represents the first ensemble classifier, and $C_2(x)$ represents the second ensemble classifier.

3.8. Hyperparameter Settings and Optimization

To enhance the reproducibility of our results, we present comprehensive details on the best hyperparameter settings for each base classifier incorporated into the proposed Concatenated Hybrid Ensemble Classifier:

Support Vector Machine (SVM):

- C: The regularization parameter managing the compromise between minimizing classification errors and optimizing the margin, with values equal to 0.1.
- gamma: The kernel coefficient in the SVM decision boundary that establishes the weight of a single training example, with values equal to 'scale'.

Decision Tree:

- max_depth: The maximum depth of the decision tree, limiting the complexity of the tree, with values equal to 5.
- min_samples_split: The minimum number of samples required to split an internal node in the tree, with values equal to 10.

K-Nearest Neighbors (KNN):

- n_neighbors: The number of nearest neighbors considered for classification, with values equal to 7.
- weights: 'Distance' (weights inversely proportional to distance) and 'Uniform' (equal weights) are the alternatives for the weight function used in prediction, with values equal to 'Uniform'.

Logistic Regression:

- C: The regularization parameter managing the trade-off between avoiding overfitting and maximizing likelihood, with values equal to one.
- penalty: The type of regularization penalty applied, with options 'l1' (L1 regularization) and 'l2' (L2 regularization), with values equal to 'l2'.

AdaBoost:

- learning_rate: A hyperparameter that scales the contribution of each weak learner. Smaller values may prevent overfitting, with values equal to 0.1.
- n_estimators: The number of weak learners (base classifiers) to combine in the ensemble, with values equal to 50.

Naive Bayes:

- var_smoothing: A smoothing parameter that adds a small value to the variances of features. This helps stabilize the computation of conditional probabilities and prevents issues with zero variances, with values equal to 1×10^{-9}.

We employed GridSearchCV, a function from the scikit-learn library, for hyperparameter optimization in fundamental classifiers: SVM, Decision Tree, KNN, Logistic Regression, AdaBoost, and Naive Bayes. Each classifier underwent a five-fold cross-validation methodology, exploring a grid of hyperparameters to identify the optimal settings.

4. Experimental Analysis

4.1. Performance Metrics

In this phase, the outcomes of the experiments have been observed and analyzed in detail. The performance of the proposed model has been evaluated using metrics such as accuracy, precision, sensitivity (recall), specificity, and F1 score.

For clarity, the definitions of these metrics are as follows:

- True Positive (TP): Correctly predicted positive cases.
- False Positive (FP): Incorrectly predicted positive cases.
- True Negative (TN): Correctly predicted negative cases.
- False Negative (FN): Incorrectly predicted negative cases.

Accuracy measures the proportion of correctly predicted instances (both positive and negative) out of the total number of instances, providing an overall measure of the classifier's correctness:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{11}$$

Precision calculates the proportion of correctly predicted positive instances out of all instances predicted as positive by the classifier. It quantifies the classifier's ability to avoid false positives:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

Sensitivity, also known as the True Positive Rate or Recall, quantifies the proportion of actual positive instances correctly identified by the classifier. It indicates the classifier's ability to identify positive instances accurately, with higher values indicating fewer false negatives:

$$Sensitivity\ (Recall) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

Specificity measures the proportion of actual negative instances correctly identified by the classifier, indicating its ability to correctly identify negative instances. A higher specificity value indicates fewer false positives:

$$Specificity = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{14}$$

The F1 score is the harmonic mean of precision and recall, providing a balance between these two metrics and considering both the classifier's ability to identify positive instances accurately and its ability to avoid false positives:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{15}$$

To facilitate model training and evaluation, we partitioned the dataset into training and validation sets. Specifically, 75% of the data were utilized for training, while the remaining 25% were allocated for validation. The standard train–test split method was employed for validation, ensuring a rigorous assessment of the proposed model's performance. This partitioning strategy enhances the robustness and generalizability of the proposed heart disease prediction model.

In the first phase of the experiment, the first ensemble classifier was built using Support Vector Machine, Decision Tree, and K-Nearest Neighbor as base classifiers. The confusion matrix generated by applying the first ensemble classifier is shown in Figure 5a. Subsequently, in the experiment, the second ensemble classifier was applied, and was constructed using Logistic Regression, Adaboost, and Naive Bayes as base classifiers. The generated confusion matrix is shown in Figure 5b. The confusion matrix of the final proposed classifier is presented in Figure 5c.

When applying the first hybrid ensemble classifier, the proposed model achieved an accuracy of 84.21%, sensitivity of 86.2%, specificity of 83.0%, precision of 75.5%, and an F1 Score of 80.6%, as shown in Table 5. Upon applying the second hybrid ensemble classifier, the model achieved an accuracy of 85.87%, sensitivity of 84.36%, specificity of 86.39%, precision of 81.85%, and an F1 Score of 83.09%. The final classifier, which was created by concatenating the two hybrid ensemble classifiers, achieved an accuracy of 86.89%, sensitivity of 87.1%, specificity of 86.7%, precision of 81.8%, and an F1 Score of 84.3%.

Table 5. Performance of hybrid and concatenated ensemble classifiers.

Classifiers	Accuracy	Sensitivity	Specificity	Precision	F1 Score
1st Ensemble	84.21%	86.2%	83.0%	75.5%	80.6%
2nd Ensemble	85.87%	84.36%	86.39%	81.85%	83.09%
Concatenated Ensemble	86.89%	87.1%	86.7%	81.8%	84.3%

In addition to previous metrics, our model's performance is comprehensively assessed using the Area Under the Receiver Operating Characteristic (AUC-ROC) curve. This metric

can offer a nuanced understanding of the model's effectiveness in healthcare predictions, considering the nuanced implications of false positives and false negatives. Figure 6 presents a visual representation of this performance metric.

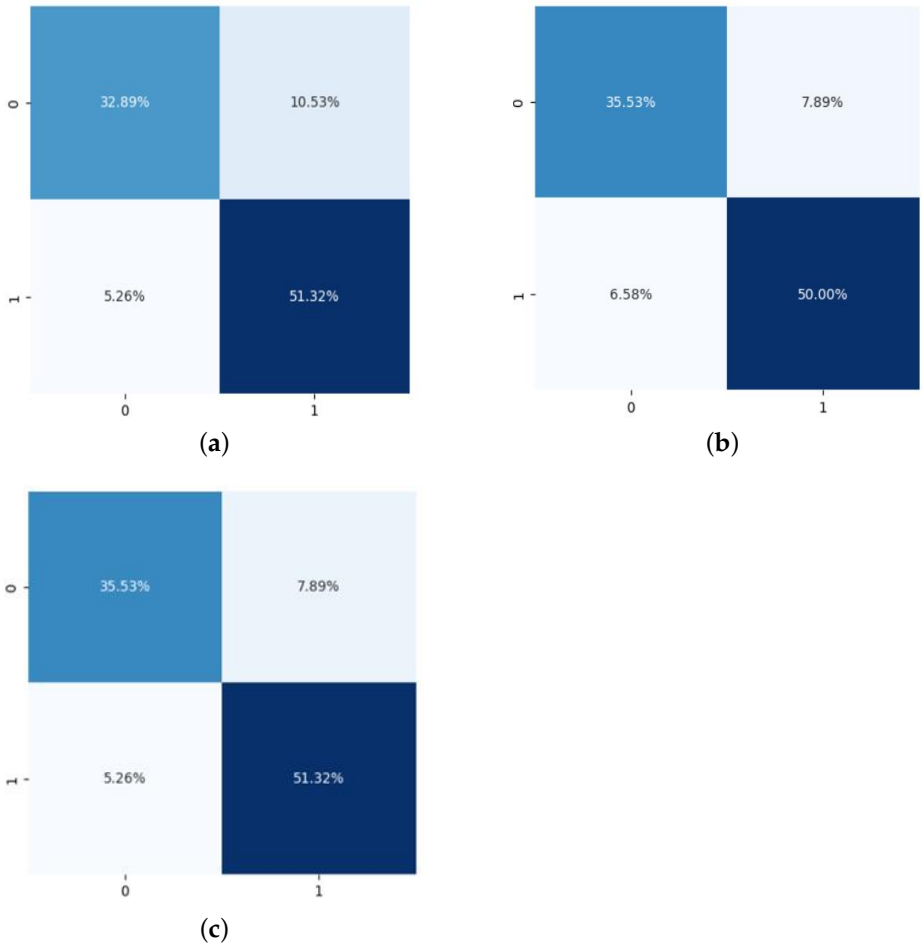


Figure 5. Distribution of age, cholesterol, oldpeak, thalach and trestbps. (**a**) Confusion matrix generated by applying Ensemble Classifier 1; (**b**) Confusion matrix generated by applying Ensemble Classifier 2; (**c**) Confusion Matrix generated by the proposed concatenated classifier.

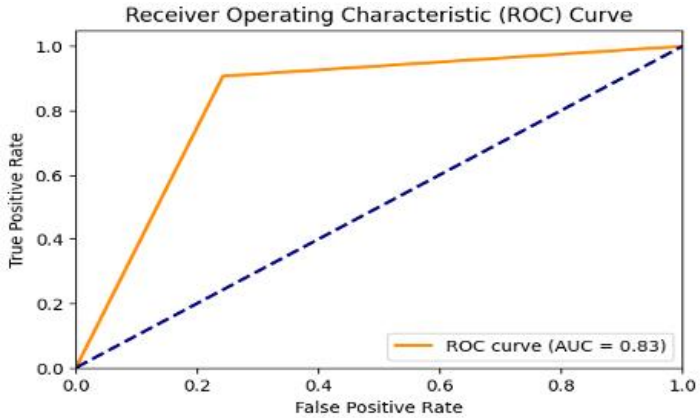


Figure 6. AUC-ROC Curve: Illustrating the model's performance in terms of True Positive Rate versus False Positive Rate.

4.2. Comparison with Existing Works

In this subsection, we present a comparison of the performance of the proposed Concatenated Hybrid Ensemble Classifier with existing works in the field of using ML algorithms for heart disease prediction. Table 6 provides a comprehensive overview of accuracy, sensitivity, specificity, precision, and F1 Score, allowing readers to assess the efficacy of our model in comparison to other relevant studies. The comparisons are based on methodologies, key observations, and performance metrics reported in related works. The proposed model's performance metrics are also highlighted to provide a clear understanding of its superiority.

Table 6. Comparison with existing works.

Existing Works	Methodology	Highest Accuracy	Highest F1 Score
[9]	Logistic Regression, Naive Bayes, Random Forest, Decision Tree	Average Accuracy: 85%	N/A
[10]	Naive Bayes, Decision Tree, KNN, Random Forest	Average Accuracy: 84%	N/A
[13]	Logistic Regression	84.53%	N/A
[15]	Semi-Supervised Self-Training	81.89%	87.14%
[20]	SVM, Gaussian Naive Bayes, Logistic Regression, LightGBM, XGBoost, Random Forest	Average Accuracy: 80%	N/A
[21]	Support Vector Machine, Decision Tree, Random Forest, Naive Bayes, Logistic Regression	N/A	N/A
[27]	Bagging Mechanism with KNN, Naive Bayes, Logistic Regression	Average Accuracy: 82%	N/A
[28]	Random Forest, Decision Tree, Logistic Regression, SVM, KNN	Average Accuracy: 75%	Highest AUC-ROC: 0.8675
Proposed Model	Concatenated Hybrid Ensemble Classifier	86.89%	84.3%

The comparison with existing works reveals several noteworthy observations. The proposed Concatenated Hybrid Ensemble Classifier consistently outperforms other models in terms of accuracy, achieving an impressive 86.89%. This accuracy surpasses the reported averages of several existing models, indicating the robustness and efficacy of our approach. Additionally, the proposed model demonstrates competitive performance in terms of F1 Score (84.3%), showcasing a balanced trade-off between precision and recall. This is crucial for applications like heart disease prediction, where both false positives and false negatives have significant implications. A deeper analysis reveals that the strength of our model lies in its ensemble strategy, which combines the strengths of diverse classifiers such as SVM, Decision Tree, KNN, Logistic Regression, AdaBoost, and Naive Bayes. This amalgamation

contributes to improved generalization and adaptability, surpassing models that rely on specific algorithms or ensemble techniques.

Despite these strengths, it is essential to acknowledge potential limitations. Our model may not excel in scenarios in which specific algorithms dominate, and further investigation into these scenarios could guide future enhancements. Practically, the superior performance of our model holds promise for accurate and reliable heart disease prediction. The combination of feature selection and data standardization enhances its predictive power and contributes to its effectiveness.

4.3. Discussion

The current research study presents a comprehensive analysis of a machine learning-based model for early heart disease prediction. The results obtained from the performed experiments highlight several key findings and implications that are crucial for understanding the effectiveness and practical utility of the model proposed in the current work. First and foremost, the model's performance metrics demonstrate its ability to accurately predict the presence or absence of heart disease. This is evident from the high accuracy, sensitivity, specificity, precision, and F1 score achieved by the proposed model. These metrics collectively indicate its effectiveness in providing reliable predictions, which is essential for early intervention and timely medical care.

One notable aspect of the proposed approach is the use of the Extra Trees Classifier for feature selection. This step was instrumental in identifying relevant features while mitigating the impact of noisy data. Feature selection is a critical component of ML models, and the use of Extra Trees contributed to the overall robustness and accuracy of the proposed model. The ensemble approach employed in our model also deserves attention. By combining multiple base classifiers, we harnessed the individual strengths of each classifier. This ensemble strategy not only improved the accuracy of the obtained predictions but also enhanced the model's ability to handle complex relationships within the data. Ensemble methods are well-suited for medical diagnosis tasks, and the derived results support their effectiveness in this context.

Interpretability is a crucial aspect of healthcare-related ML models. The proposed ensemble of classifiers provides a level of interpretability by allowing us to analyze the contributions of each base classifier to the final prediction. This interpretability can aid medical professionals in understanding the model's decision-making process and gaining insights into the factors that influence predictions. Generalizability, a key requirement for practical applicability, was observed in our model. It consistently performed well on both training and testing datasets, indicating its ability to make reliable predictions on unseen patient data. This is a vital characteristic for any model intended for clinical use. From a clinical perspective, the model's capability for early heart disease detection holds significant promise. Early detection can lead to timely intervention and personalized patient care, potentially reducing the mortality rates associated with heart disease.

Looking ahead, there are several avenues for future research. Further exploration of feature engineering techniques, the incorporation of more diverse and extensive datasets, and the integration of advanced deep learning models could enhance the model's performance. Additionally, rigorous validation on a larger and more diverse patient population is essential to establish its real-world clinical utility. In conclusion, our study presents a robust and interpretable machine learning model for heart disease prediction. The combination of feature selection, ensemble learning, and effective base classifiers contributes to its success. With further refinement and validation, this model has the potential to be a valuable tool for early heart disease detection, ultimately improving patient care and outcomes.

5. Conclusions and Future Work

In this paper, we have presented a heart disease prediction model that leverages feature selection, data standardization, and a concatenated hybrid ensemble voting classifier. The results of the performed experiments demonstrate the model's promising capability

to accurately predict heart disease. By utilizing the Extra Trees Classifier for feature selection and StandardScaler for data standardization, we have enhanced the model's overall performance and reliability.

The standout feature of our approach is the concatenated ensemble classifier, which combines the strengths of multiple base classifiers. This amalgamation results in improved accuracy, robustness, and interpretability of the model. These findings underscore the potential of machine learning techniques in advancing heart disease prediction and aiding clinical decision-making and patient care.

There is certainly ample room for future research in the domain of the current research. One avenue is the exploration of more sophisticated feature engineering techniques which could be used to further refine the model's predictive capabilities. Additionally, the incorporation of larger and more diverse datasets from varied demographics could enhance the model's generalization and real-world applicability. Furthermore, deep learning models and neural networks warrant investigation as potential additions to our approach, potentially improving prediction accuracy. Rigorous validation on a broader patient population is essential to establish the model's clinical utility and efficacy.

In conclusion, this work contributes to the ongoing efforts in the development of heart disease prediction models based on ensemble machine learning methods. Our proposed model shows promise and opens up exciting possibilities for future research in the field of using Artificial Intelligence for cardiovascular health care. Ultimately, the advancements in this research field hold the potential to positively impact clinical practices and patient outcomes.

Author Contributions: Conceptualization, A.B.M., S.G., D.S., B.A., V.C.G., A.K. and P.P.; Data Curation, A.B.M., S.G., D.S., B.A., V.C.G., A.K. and P.P.; Writing—Original Draft, A.B.M., S.G., D.S., B.A., V.C.G., A.K. and P.P.; Methodology, A.B.M., S.G., D.S., B.A., V.C.G., A.K. and P.P.; Review and Editing, A.B.M., S.G., D.S., B.A., V.C.G., A.K. and P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

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

References

1. Celermajer, D.S.; Chow, C.K.; Marijon, E.; Anstey, N.M.; Woo, K.S. Cardiovascular Disease in the Developing World: Prevalences, Patterns, and the Potential of Early Disease Detection. *J. Am. Coll. Cardiol.* **2012**, *60*, 1207–1216. [CrossRef]
2. Bakar, W.A.W.A.; Josdi, N.L.N.B.; Man, M.B.; Zuhairi, M.A.B. A Review: Heart Disease Prediction in Machine Learning & Deep Learning. In Proceedings of the 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Kedah, Malaysia, 3–4 March 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 150–155.
3. Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms* **2023**, *16*, 88. [CrossRef]
4. Cardiovascular Diseases (CVDs). Available online: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed on 10 November 2023).
5. Jánosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. Heart Disease. *Uci Mach. Learn. Repos.* **1988**. [CrossRef]
6. Désir, C.; Petitjean, C.; Heutte, L.; Salaün, M.; Thiberville, L. Classification of Endomicroscopic Images of the Lung Based on Random Subwindows and Extra-Trees. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 2677–2683. [CrossRef] [PubMed]
7. Raju, V.N.G.; Lakshmi, K.P.; Jain, V.M.; Kalidindi, A.; Padma, V. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. In Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 729–735.
8. Dietterich, T.G. Ensemble Methods in Machine Learning. In Proceedings of the 1st International Workshop on Multiple Classifier Systems (MCS), Nanjing, China, 15–17 May 2000; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany; Volume 1857, pp. 1–15.
9. Rajdhan, A.; Agarwal, A.; Sai, M.; Ravi, D.; Ghuli, P. Heart Disease Prediction using Machine Learning. *Int. J. Eng. Res. Technol. (IJERT)* **2020**, *9*, 440–450.
10. Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* **2020**, *1*, 345. [CrossRef]

11. Haq, A.U.; Li, J.; Memon, M.H.; Nazir, S.; Sun, R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mob. Inf. Syst.* **2018**, *2018*, 3860146. [CrossRef]
12. Swain, D.; Parmar, B.; Shah, H.; Gandhi, A.; Pradhan, M.R.; Kaur, H.; Acharya, B. Cardiovascular Disease Prediction using Various Machine Learning Algorithms. *J. Comput. Sci.* **2022**, *18*, 993–1004. [CrossRef]
13. Mridha, K.; Kuri, A.C.; Saha, T.; Jadeja, N.; Shukla, M.; Acharya, B. Toward Explainable Cardiovascular Disease Diagnosis: A Machine Learning Approach. In Proceedings of the International Conference on Data Analytics and Insights (ICDAI), Kolkata, India, 11–13 May 2023; pp. 409–419.
14. Triguero, I.; García, S.; Herrera, F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowl. Inf. Syst.* **2015**, *42*, 245–284. [CrossRef]
15. Livieris, I.E.; Kanavos, A.; Tampakas, V.; Pintelas, P.E. An Auto-Adjustable Semi-Supervised Self-Training Algorithm. *Algorithms* **2018**, *11*, 139. [CrossRef]
16. Boukhatem, C.; Youssef, H.Y.; Nassif, A.B. Heart Disease Prediction Using Machine Learning. In Proceedings of the Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 21–24 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
17. Sharma, V.; Yadav, S.; Gupta, M. Heart Disease Prediction using Machine Learning Techniques. In Proceedings of the 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 18–19 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 177–181.
18. 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]
19. Patel, A.C.; Shameem, A.; Chaurasiya, S.; Mishra, M.; Saxena, A. Prediction of Heart Disease Using Machine Learning. *Int. J. Sci. Dev. Res.* **2019**, *4*, 354–357
20. Karthick, K.; Aruna, S.K.; Samikannu, R.; Kuppusamy, R.; Teekaraman, Y.; Thelkar, A.R. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. *Comput. Math. Methods Med.* **2022**, *2022*, 6517716. [CrossRef]
21. Divya, K.; Sirohi, A.; Pande, S.; Malik, R. An IoMT Assisted Heart Disease Diagnostic System Using Machine Learning Techniques. In *Cognitive Internet of Medical Things for Smart Healthcare*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 145–161.
22. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554. [CrossRef]
23. Ramprakash, P.; Sarumathi, R.; Mowriya, R.; Nithyavishnupriya, S. Heart Disease Prediction Using Deep Neural Network. In Proceedings of the International Conference on Inventive Computation Technologies (ICICT), Nepal, Lalitpur, India, 26–28 February 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 666–670.
24. Pasha, S.N.; Ramesh, D.; Mohmmad, S.; Harshavardhan, A.; Shabana. Cardiovascular disease prediction using deep learning techniques. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2020; Volume 981, p. 022006.
25. Grampurohit, S.; Sagarnal, C. Disease Prediction using Machine Learning Algorithms. In Proceedings of the International Conference for Emerging Technology (INCET), Belgaum, Karnataka India, 5–7 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–7.
26. Vanitha, G.; Shalini, K.; Shivani, C. Heart Disease Prediction Using Hybrid Technique. *J. Interdiscip. Cycle Res.* **2020**, *6*, 920–927.
27. Majumder, A.B.; Gupta, S.; Singh, D. An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2022; Volume 2286, p. 012017.
28. Kumar, N.K.; Sindhu, G.S.; Prashanthi, D.K.; Sulthana, A.S. Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers. In Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 15–21.
29. Kota, P.; Madenahalli, A.; Guturi, R.; Nukala, B.; Nagaraj, S.; Kota, S.; Neeli, P.C. Heart Disease Classification Comparison among Patients and Normal Subjects using Machine Learning and Artificial Neural Network Techniques. *Int. J. Biosens. Bioelectron.* **2021**, *7*. [CrossRef]
30. Swarupa, A.N.V.K.; Sree, V.H.; Nookambika, S.; Kishore, Y.K.S.; Teja, U.R. Disease Prediction: Smart Disease Prediction System using Random Forest Algorithm. In Proceedings of the International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, 13–14 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 48–51.
31. Islam, S.; Jahan, N.; Khatun, M.E. Cardiovascular Disease Forecast using Machine Learning Paradigms. In Proceedings of the 4th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 11–13 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 487–490.
32. Livieris, I.E.; Kanavos, A.; Tampakas, V.; Pintelas, P.E. An Ensemble SSL Algorithm for Efficient Chest X-ray Image Classification. *J. Imaging* **2018**, *4*, 95. [CrossRef]
33. Livieris, I.E.; Kiriakidou, N.; Kanavos, A.; Tampakas, V.; Pintelas, P.E. On Ensemble SSL Algorithms for Credit Scoring Problem. *Informatics* **2018**, *5*, 40. [CrossRef]
34. UCI Heart Disease Data. Available online: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data (accessed on 10 November 2023).
35. Baeza-Yates, R.A.; Ribeiro-Neto, B.A. *Modern Information Retrieval*; ACM Press: New York, NY, USA; Addison-Wesley: Boston, MA, USA, 1999.

36. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [CrossRef]
37. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
38. Song, Y.Y.; Lu, Y. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130.
39. Peterson, L.E. K-Nearest Neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
40. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2000.
41. Menard, S. *Applied Logistic Regression Analysis*; Number 106; Sage: Thousand Oaks, CA, USA, 2002.
42. Schapire, R.E. Explaining AdaBoost. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 37–52.
43. Rish, I. An Empirical Study of the Naive Bayes Classifier. In Proceedings of the IJCAI Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4–6 August 2001; Volume 3; pp. 41–46.