Journal of Data Analysis and Information Processing, 2025, 13(2), 125-139
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Extra Trees Model for Heart Disease Prediction

Uchenna J. Nzenwata[1*], Emokiniovo Edwin[1], Emmanuel A. Chukwu[2], Dare Osilaja[3], Johnson O. Hinmikaiye[1], Chidiebere Enyinnah[1]

[1]Computer Science Department, Babcock University, Ilishan-Remo, Ogun State, Nigeria
[2]Computer Science Department, Adeleke University, Ede, Osun State, Nigeria
[3]Department of Applied Mathematics, Wrocław University of Science and Technology, Wrocław, Poland
Email: *nzenwatau@babcock.edu.ng

How to cite this paper: Nzenwata, U.J., Edwin, E., Chukwu, E.A., Osilaja, D., Hinmikaiye, J.O. and Enyinnah, C. (2025) Extra Trees Model for Heart Disease Prediction. *Journal of Data Analysis and Information Processing*, **13**, 125-139.
https://doi.org/10.4236/jdaip.2025.132008

Received: March 7, 2025
Accepted: April 27, 2025
Published: April 30, 2025

Abstract

Heart disease continues to be a major global cause of death, making the development of reliable prediction models necessary to enable early detection and treatment. Using machine learning to improve prediction accuracy, this study investigates the use of the Extra Tree (Extremely Randomized Trees) algorithm for heart disease prediction. The research includes data preparation, model training, and performance evaluation using measures like accuracy, precision, recall, and F1-score. It makes use of a dataset that includes a variety of medical and demographic variables. The Extra Tree model outperforms a number of baseline models in terms of accuracy and predictive power. The dataset was obtained from the University of California, Irvine (UCI) Machine Learning Repository, which contains about 319,796 instances and 18 attributes related to heart disease. The attributes serve as the features. This study reduced the number of features from 18 to 7, by using recursive feature elimination method, which uses Random Forest as an estimator. The Extra Tree model demonstrates great performance, showing high accuracy, precision, recall, and f1 scores of 93.1%, 94.8%, 100% and 93.1% respectively on a dataset split ratio of 80% to 20% train set and test set respectively. The study concluded that the model may be implemented into a clinical decision support system to help healthcare providers diagnose cardiac disease. Furthermore, the feature importance analysis can help direct future research into finding the most significant risk factors for cardiovascular disease.

Keywords

Accuracy, Extra tree Model, Heart Disease Prediction, Machine Learning, Predictive Model, Random Forest, Recursive Feature Elimination

1. Introduction

Heart disease is a major global health concern and the world's top cause of mortality. The World Health Organisation estimates that 17.9 million fatalities worldwide in 2016 were related to cardiovascular illnesses, or 31% of all deaths [1]. Heart disease mortality and major complications can be greatly decreased by early detection and prevention of the condition. It is impossible to overestimate the importance of correctly forecasting cardiac disease since making a mistake might have grave repercussions. 17.9 million deaths globally are attributed to heart disease each year [2]. Numerous unhealthy behaviours, including obesity, hypertension, and excessive cholesterol, are to blame for the growth of heart disease [3]. However, it might be difficult to diagnose heart disease correctly since its symptoms often resemble those of other illnesses [4].

Machine learning has become increasingly important in the medical field, with the potential to improve the diagnosis and prediction of diseases [5]. Studies have shown that machine learning models can be used to classify and predict heart disease diagnosis, with some achieving accuracy rates of up to 94.60% [6]. Machine learning algorithms such as CART [7], deep neural networks [8], random forest [9], and support vector machines [10] have been used to detect heart disease. Despite the promising results, there are challenges associated with using machine learning for heart disease prediction, which is associated with high data dimensionality. High dimensionality of data can result in overfitting, requiring the use of feature engineering and selection techniques to reduce data redundancy and processing time [11]. Dimensionality reduction techniques such as PCA can also be used to store valuable information in new components [12].

Several studies [13]-[15] have been conducted using the heart disease 2020 project dataset, a popular benchmark dataset for heart disease prediction [16]. Previous studies used K-Nearest neighbours, Random Forest, Decision Tree, Support Vector Machines, and extreme gradient Boosting classifiers. The studies have achieved high accuracy rates, including random forest with 83.2% accuracy, decision tree with 86.1% accuracy, and SVM with 82.7% accuracy, among others. However, there is still room for improvement, particularly in terms of reducing the dimensionality of the dataset and using appropriate classification techniques. Therefore, this study adopted the use of random forest as the features selection estimator, and Extra Tree machine learning classifier for the prediction of heart disease.

Research Problem and Objective of the Study

Heart disease is a significant global health issue, being the leading cause of death worldwide. Early and accurate prediction of heart disease is crucial to prevent severe complications and fatalities. Machine learning algorithms have shown promising results in predicting heart disease, but there are still challenges to be addressed. The main problem is the high dimensionality of the data used in these algorithms, which can lead to overfitting and require large amounts of memory

for processing. Previous studies have used machine learning models such as Logistic Regression (LR), K-Nearest Neighbours (KNN), Decision Trees (DT), Extreme Gradient Boosting (XGBoost), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF) to predict heart disease using the heart disease 2020 project dataset. The outcome of these models yielded 91%, 93.3%, 86.5%, 91.4%, 84.7%, 91.5%, and 90.5% accuracy scores respectively. This implies that there is room for improvement as no study has been able to use the Extra Tree machine learning classifier. Therefore, this study develops a machine learning model using the Extra Tree classifier to predict heart disease in the face of the heart disease 2020 project dataset.

The findings of this study are expected to contribute to the field of healthcare by providing a reliable and accurate method for predicting heart disease and also making heart disease diagnosis available to the public. This can aid healthcare professionals in early diagnosis and treatment planning, ultimately improving patient outcomes.

2. Review of Literature

A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction was carried out in [17]. The authors aimed to develop a machine learning algorithm for accurate prediction and decision making for cardiovascular disease (CVD) patients. Machine learning algorithms (DT, RF, LR, NB, SVM) were implemented for classification and prediction. The random forest (RF) algorithm had the highest accuracy, sensitivity, and recursive operative characteristic curve for CVD classification and prediction. However, the study did not show the sample size or representativeness of the study population and there was a lack of discussion on potential biases or limitations of the machine learning algorithms used. Another study was done by [18], where predictive modeling of cardiovascular disease using Machine Learning Techniques. Random forest, decision tree, gradient-boosted tree, logistic regression. Gradient-boost tree algorithm obtained more accurate results than other methods.

In an attempt to enhance heart disease prediction through ensemble learning techniques with hyperparameter optimization Algorithms, the outcome of the study in [19] looked promising, but the authors did not demonstrate a realistic approach that led to the development of the ensemble model. The study in [20] predicted cardiovascular disease using feature selection techniques. The study used six different supervised machine learning techniques for analysis. The accuracy, precision, recall, and f-measure were compared for each classifier. The study compares the accuracy, precision, recall, and f-measure of six different machine learning techniques. Logistic regression and SVM have better accuracy than other classifiers with 89.6% and 89.8% accuracy scores respectively.

A study developed an ensemble model in [21] through the meta-analytic approach that assesses the predictive ability of machine learning algorithms in cardiovascular diseases dataset. Comprehensive search strategy executed within

MEDLINE, Embase, and Scopus databases was done. The study included 103 cohorts with a total of 3,377,318 individuals. The study shows that there is potential in ensemble models to predict heart disease. The accuracy score of the best model was benchmarked on 90% accuracy score. The study of [22] aimed to predict death caused by cardiovascular disease within ten years of follow-up. The machine learning used are the Logistic Regression, Support Vector Machine, Random Forest, Naïve Bayes, Extreme Grading Boosting, and Adaptive Boosting. The evaluation metrics used are Accuracy, Precision, Recall, F1-Score, Specificity, and AUC. The Utilized study used clinical CVD risk factors and biochemical data, and the outcome shows that the Logistic Regression had the most reliable algorithm assessment with 72.20% accuracy.

In implementing an early identification and treatment outcomes for cardiovascular disease, the authors in [23] made use of neural networks, random forests, Bayesian networks, C5.0, and QUEST on a dataset obtained from Kaggle with 70,000 patient records used for training and validation. An accurate score of 93% was recorded to show a promising sign for early detection. The study in [24] uses a fundamental model capable of performing various heart disease analysis algorithms. The model caters to a specific type of data. Optimization techniques were employed for data classification and prediction accuracy enhancement. The outcome provided a predictive framework for various heart diseases using decision tree, random forest, logistic regression, and KNN algorithms. The framework offers an interface that's easy to use.

In the quest to improve the predicting accuracy of the existing model, the study in [25] developed a model using k-modes clustering with Huang starting that can improve classification accuracy. Models utilized include random forest (RF), decision tree classifier (DT), multilayer perceptron (MP), and XGBoost (XGB). GridSearchCV was used to hypertune the parameters of the applied model in order to improve the outcome. The suggested model is tested on a real-world dataset of 70,000 occurrences from Kaggle. Models were trained on data with an 80:20 split and achieved the following accuracy: Decision tree: 86.37% (with cross-validation) and 86.53% (without cross-validation); XGBoost: 86.87% (with cross-validation) and 87.02% (without cross-validation); random forest: 87.05% (with cross-validation) and 86.92% (without cross-validation); multilayer perceptron: 87.28% (with cross-validation) and 86.94% (without cross-validation). The proposed models have the following AUC (area under the curve) values: Decision tree: 0.94, XGBoost: 0.95, Random Forest: 0.95, and Multilayer Perceptron: 0.95. The conclusion drawn from this underlying research is that multilayer perceptron with cross-validation has outperformed all other algorithms in terms of accuracy. It achieved the highest accuracy of 87.28%. The study in [26] provides a thorough analysis of the several machine learning approaches that are available and examines how well they work for effective heart disease diagnosis, treatment, and prediction. The suggested research surveys various machine learning techniques, such as support vector machines (SVM), decision trees (DT), Naïve Bayes (NB),

K-nearest neighbor (KNN), artificial neural networks (ANN), etc., that are employed to forecast the development of heart illnesses. The best and worst performing techniques overall were then determined by calculating the average forecast accuracy for each technique. The results showed that the C4.5 decision tree technique produced the lowest average prediction accuracy of 74.0%, while the ANN achieved the best average prediction accuracy of 86.91%.

From the reviewed studies, it was discovered that the existing studies have good accuracy scores on the prediction of heart diseases. However, there is room for improvement as some of the reviews are weak in terms of feature selection approaches and improper tuning of the learning models. Also, it appears that there are not many studies that have employed the use of Extra Tree Classifier. Therefore, this current study focuses on how features can be selected using the recursive feature elimination through the use of Random Forest estimator, and the development of a predictive model using the extra tree classifier.

2.1. Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) is a broad field that encompasses the development of intelligent machines that can perform tasks that typically require human intelligence. AI systems can mimic human cognitive functions such as learning and problem-solving through the use of mathematics and logic [27].

Machine Learning (ML), on the other hand, is a subset of AI. It is an application of AI that enables a computer system to learn without direct instruction. This is achieved by using mathematical models of data to help the computer learn and improve on its own, based on experience [27]. The connection between AI and ML is that an "intelligent" computer uses AI to think like a human and perform tasks on its own. Machine learning is how a computer system develops its intelligence. The process of creating an AI system involves building it using machine learning and other techniques. Machine learning models are created by studying patterns in the data, and data scientists optimize these models based on patterns in the data. This process repeats and is refined until the models' accuracy is high enough for the tasks that need to be done.

AI and ML have numerous capabilities that have become valuable in helping companies transform their processes and products. Some of these capabilities include predictive analytics, recommendation engines, speech recognition, natural language understanding, image and video processing, and sentiment analysis [27]. The connection between AI and ML offers powerful benefits for companies in almost every industry. Some of the top benefits that companies have already seen include improved efficiency, better decision-making, and the ability to create new products and services.

2.2. Machine Learning Algorithms

This section identifies and discusses the relevant machine learning algorithms that were considered in this study.

1) Logistic Regression (LR)

Logistic Regression is a statistical model used for binary classification problems. It uses a logistic function to model the probability of a binary response based on one or more predictor variables. The goal of logistic regression is to find the best fitting model to describe the data and to use it for prediction [27].

2) K-Nearest Neighbours (KNN)

KNN is a type of instance-based learning algorithm that can be used for both classification and regression problems. It works by finding the k-nearest data points in the feature space to the new data point, and then assigning the new data point to the class that is most common among its k-nearest neighbours [27].

3) Decision Trees (DT)

Decision Trees are a type of supervised learning algorithm that is mostly used in classification problems. It works by partitioning the feature space into a series of rectangular regions, with each region corresponding to a leaf node in the tree. The goal of a decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [27].

4) Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting algorithm that is used for both regression and classification problems. It works by building an ensemble of weak prediction models, typically decision trees, in a stage-wise fashion. It generalizes them by allowing optimization of an arbitrary differentiable loss function [27].

5) Naïve Bayes (NB)

Naïve Bayes is a family of probabilistic classification algorithms based on Bayes' theorem. It is called "naïve" because it makes the "naïve" assumption that the features are conditionally independent given the class. Despite this assumption, Naïve Bayes often works well in practice [27].

6) Support Vector Machines (SVM)

SVM is a supervised learning algorithm used for both classification and regression problems. It works by finding the hyperplane that maximally separates the data points of different classes. In cases where no hyperplane exists, SVM uses the kernel trick to transform the data into a higher dimensional space where a separating hyperplane can be found [27].

7) Random Forest (RF)

Random Forest is an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forest corrects for decision trees' habit of overfitting to their training set by building a multitude of trees and outputting the class that is the mode of the classes or mean prediction of the individual trees [27].

8) Extra Tree Classifier (ETC)

An Extra Tree Classifier is a type of decision tree ensemble algorithm that is used for classification tasks. It differs from a classic decision tree in the way it selects

the split at each node. Instead of choosing the best split among all possible splits, Extra Tree Classifier selects the split at random. This randomness helps to reduce overfitting and improves the robustness of the model [28]. The key parameters of the Extra Tree Classifier are:

- **criterion:** The function to measure the quality of a split. The supported criteria are "gini" for the Gini impurity and "entropy" for the Shannon information gain.
- **splitter:** The strategy used to choose the split at each node. The supported strategies are "best" to choose the best split and "random" to choose the best random split.
- **max_depth:** The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- **min_samples_split:** The minimum number of samples required to split an internal node.
- **min_samples_leaf:** The minimum number of samples required to be at a leaf node.
- **max_features:** The number of features to consider when looking for the best split.
- **random_state:** Used to pick randomly the max_features used at each split.

The Extra Tree Classifier also has an attribute feature_importances that can be used to determine the importance of each feature in the classification task.

3. Methodology

The dataset used in this study was obtained from the University of California, Irvine (UCI) Machine Learning Repository. The dataset contains over 300 thousand patient records, with 18 features (with 17 predictors and 1 target). The dataset is text data in the comma separated value (csv) format. The dataset in its original form contains some irregularities that were handled through data preprocessing stages to get it to a machine trainable format. The data features were normalized or standardized to ensure that each feature had equal weight in the analysis. This was crucial to ensure the machine learning algorithms' prediction accuracy and reliability. Relevant features that contributed to the accurate prediction of heart disease were selected using Recursive Feature Elimination (RFE). The selected features were used to train the Extra Tree Classifier (ETC) Model. The Extra Tree model was implemented using a Python programming language in the Google Colab environment. Thereafter, a streamlit Python framework was used to develop an interface for the model. The developed model's performance was evaluated using evaluation metrics such as accuracy, precision, recall, and F1-score.

3.1. Experimental Set-Up

The data features were transformed by using a label encoder scaler transformer. This was used to transform the categorical data features and to scale the feature

values to a standard acceptable to the training algorithms.

Figure 1 describes the summary of the data preprocessing phase.

The feature selection was done to ensure that possible model over/underfitting is avoided. Also, this helps to ensure that features of high importance are used. This study used the recursive feature elimination (RFE) method. For a proper estimation and random selection of the features, the random forest was used as the estimator. Figure 2 shows the workflow of the RFE.

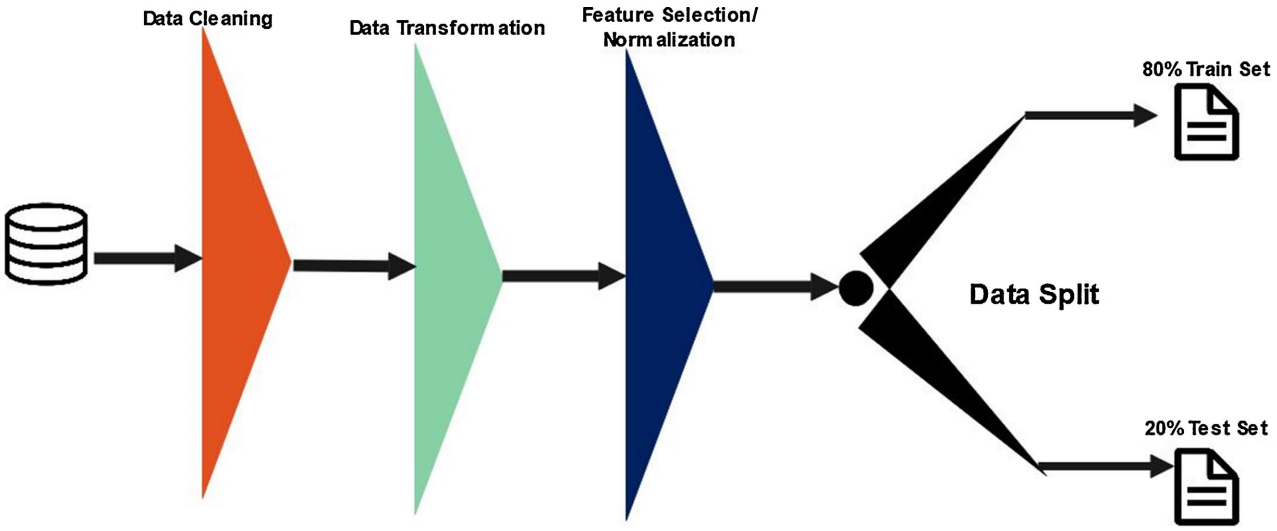

Figure 1. Data Preprocessing workflow.

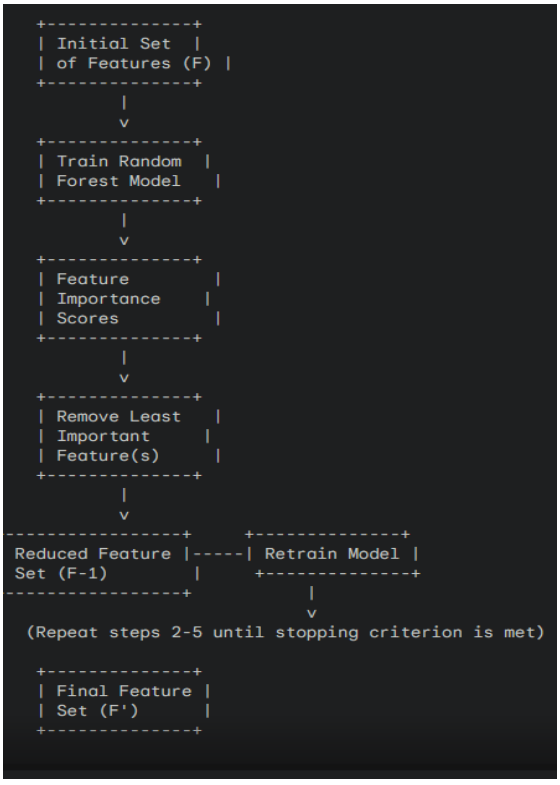

Figure 2. RFE feature selection model.

RFE is a wrapper method for feature selection that utilizes an estimator (like a Random Forest) to iteratively remove features deemed least important. Here's a breakdown of the process as shown in Figure 2:

1) Start with All Features: Begin with the entire set of features in your dataset.

2) Train a Random Forest: Create a Random Forest model and train it on the complete dataset.

3) Evaluate Feature Importance: Extract feature importance scores from the Random Forest. These scores indicate how much each feature contributes to the model's predictions.

4) Remove Least Important Feature(s): Identify the feature with the lowest importance (or a subset of features if you're removing multiple at once). Eliminate this feature (or features) from the dataset.

5) Repeat: Go back to step 2 and retrain the Random Forest on the reduced dataset.

6) Continue Recursively: Repeat steps 2 - 5 until you reach a desired number of features, a performance metric threshold, or another stopping criterion.

The model was designed using the extra tree classifier. The trained ETC was compared with existing models. Figure 3 shows the general framework of the model for the heart disease prediction.

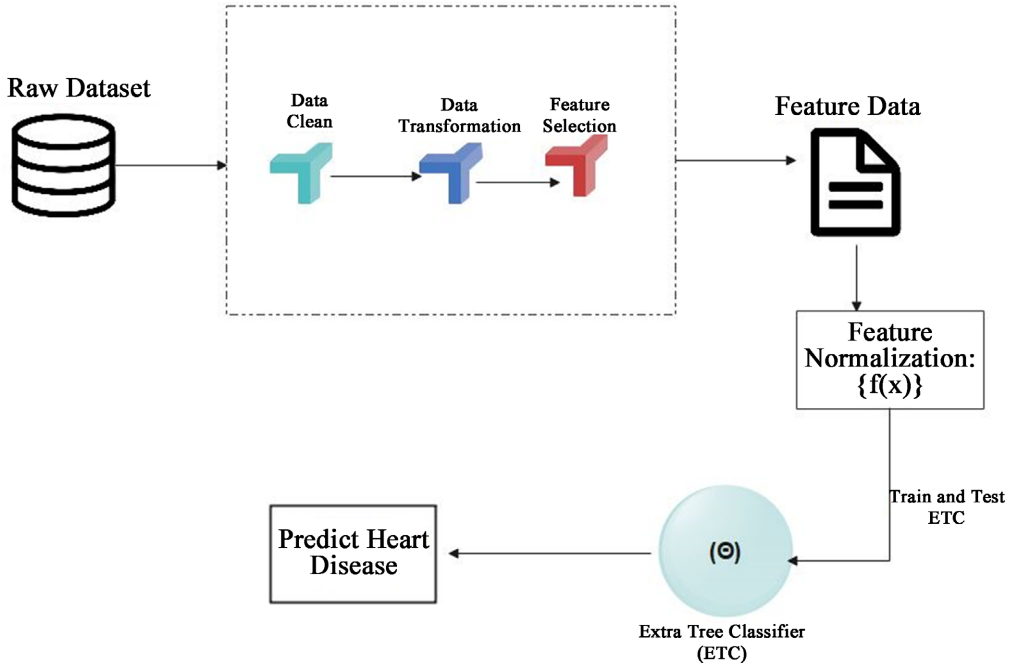

Figure 3. The general heart disease prediction framework.

3.2. Perform Feature Selection Using Recursive Feature Elimination (RFE)

The data features were selected and reduced. This was done to actually select the features that contribute more to the prediction of heart disease. The study made use of RFE method using Random Forest as the estimator. Figure 4 shows the

features that were selected after applying feature elimination method. The method ranked 10 features higher than others by assigning value = 1 to the highly ranked features.

```
RandomForest Feature Rankings:
              Feature  Rank
0                 BMI     1
1             Smoking     5
2     AlcoholDrinking     8
3              Stroke     1
4      PhysicalHealth     1
5        MentalHealth     1
6         DiffWalking     1
7                 Sex     3
8         AgeCategory     1
9                Race     1
10           Diabetic     1
11  PhysicalActivity     2
12          GenHealth     1
13          SleepTime     1
14             Asthma     4
15      KidneyDisease     7
16        SkinCancer     6
```

Figure 4. RFE ranked features.

The selected features are: ["BMI", "Stroke", "PhysicalHealth", "MentalHealth", "AgeCategory", "GenHealth", "SleepTime"]. These features will be used to train the ExtraTree model.

3.3. Model Development

In developing the model, this study implemented the existing models as discussed in the body of literature to verify the obtained results. The models that were verified are: Logistic Regression (LR), K-Nearest Neighbours (KNN), Decision Trees (DT), Extreme Gradient Boosting (XGBoost), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF).

Table 1 shows the models and their parameters used for the training.

Table 1. Models and set parameters.

Models	Parameters & Values
Logistic Regression (LR)	solver = "liblinear", C = 0.1, penalty = "l1"
K-Nearest Neighbours (KNN)	n_neighbors = 5, metric = "euclidean"
Decision Trees (DT)	criterion = "gini", max_depth = 3, min_samples_split = 2, min_samples_leaf = 1
Extreme Gradient Boosting (XGBoost)	objective = "binary: logistic", eval_metric = "auc", max_depth = 5, learning_rate = 0.1, n_estimators = 100
Naïve Bayes (NB)	var_smoothing = 1.0
Support Vector Machines (SVM)	kernel = "rbf", C = 1.0, gamma = 0.1
Random Forest (RF)	n_estimators = 100, max_depth = 5, min_samples_split = 2, min_samples_leaf = 1
Our Model (ExtraTree Classifier)	n_estimators = 100, max_depth = 5, min_samples_split = 2, min_samples_leaf = 1, oob_score = True

4. Result and Discussion

The result obtained was compared with the results of the individual models. The basis of comparison was dependent on the use of similar dataset, parameter tuning and the evaluation metrics. The evaluation metrics were used to determine the extent of the correctness of the predictions.

Table 2 shows the summary of the results of this study. From the results, it was discovered that the Extra Tree model outperformed other models with the accuracy, precision, recall and f1-scores of 93.1%, 94.8%, 100%, and 93.1% respectively.

Table 2. Result summary.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	91.3	92.0	99.0	96.0
K-Nearest Neighbour	90.4	92.0	98.0	95.0
Decision Tree	86.3	83.4	89.0	88.0
Extreme Gradient Boost (XGBoost)	91.2	85.0	83.0	81.0
Naïve Bayes	84.3	82.0	86.0	83.0
Random Forest	91.3	93.2	1	98.1
Support Vector Machine	91.5	92.0	1	98.1
Extra Tree	93.1	94.8	1	98.9

The Extra Tree models, also called extremely randomized trees, show superiority because they offer a good balance between bias and variance, making them effective for various machine learning tasks. The tabulated result was represented in a graphical form for a proper result visualization. Figure 5 shows the graphical presentation of the result.

The Extra model was adopted for the prediction of heart disease. Figure 6

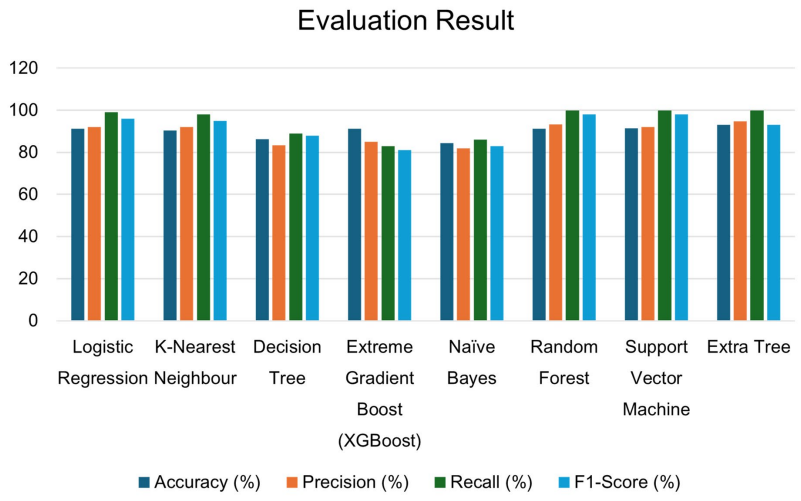

Figure 5. The graphical presentation of the obtained result.

Actual Values	Predicted Values
0	0
0	0
0	0
0	0
0	0
1	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0

Figure 6. Prediction output.

shows the output of the first 20 predictions. The actual values are the values presented in the test dataset, while the predicted values are the values predicted using the extra tree model. The possibility of having heart disease is encoded using "1" and the possibility of no heart disease is encoded using "0".

The Extra model was adopted for the prediction of heart disease. Figure 6 shows the output of the first 20 predictions. The actual values are the values presented in the test dataset, while the predicted values are the values predicted using the extra tree model. The possibility of having heart disease is encoded using "1" and the possibility of no heart disease is encoded using "0".

The evaluation metrics used for the models' calibration are the accuracy, precision, recall and f1-score. The outcome of this study shows that the Extra tress model has the highest performance scores in terms of prediction when the results are compared with the other popular machine learning models to benchmark the effectiveness of the Extra Tree model. The Extra Tree model demonstrates great performance, showing high accuracy, precision, recall, and f1 scores of 93.1%, 94.8%, 100% and 98.9% respectively on an imbalanced dataset split ratio of 80% to 20% train set and test set respectively.

5. Conclusion

The study comes to the conclusion that the Extra Tree model is a very useful tool for heart disease prediction. The model is especially well-suited for medical prediction tasks because of its resilience against overfitting and capacity to handle big datasets with many characteristics. The Extra Tree model may be effectively employed in clinical settings to support healthcare professionals in the early identification and prevention of heart disease, as demonstrated by its high accuracy and

robust performance metrics. This study's findings suggest that the Extra Trees model should be utilized to predict heart disease. The model may be implemented into a clinical decision support system to help healthcare providers diagnose cardiac disease. Furthermore, the feature importance analysis can help direct future research into finding the most significant risk factors for cardiovascular disease. It is also advised that the model be validated on bigger and more varied datasets to confirm its generalizability and robustness.

Ethical Consideration

This study will be guided by the principle and code of practice of the Babcock University Health Research Ethics Committee (BUHREC), and will be conducted under thorough supervision by experts in the related areas of specialization. As a result, it guarantees that the intellectual property rights of the secondary dataset used, machine learning algorithms, programming tools, and works of literature included in this study's documentation are correctly referenced in accordance with the Institute of Electrical and Electronics Engineers (IEEE) ethical standard.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mukasheva, G., Abenova, M., Shaltynov, A., Tsigengage, O., Mussabekova, Z., Bulegenov, T., *et al.* (2022) Incidence and Mortality of Cardiovascular Disease in the Republic of Kazakhstan: 2004-2017. *Iranian Journal of Public Health*, **51**, 821-830. https://doi.org/10.18502/ijph.v51i4.9243

[2] Dai, H., Bragazzi, N.L., Younis, A., Zhong, W., Liu, X., Wu, J., *et al.* (2021) Worldwide Trends in Prevalence, Mortality, and Disability-Adjusted Life Years for Hypertensive Heart Disease from 1990 to 2017. *Hypertension*, **77**, 1223-1233. https://doi.org/10.1161/hypertensionaha.120.16483

[3] Upadhyay, R.K. (2022) Chronic Non-Communicable Diseases: Risk Factors, Disease Burden, Mortalities and Control. *Acta Scientific Medical Sciences*, **6**, 153-170. https://doi.org/10.31080/asms.2022.06.1227

[4] Ahsan, M.M. and Siddique, Z. (2022) Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review. *Artificial Intelligence in Medicine*, **128**, Article ID: 102289. https://doi.org/10.1016/j.artmed.2022.102289

[5] Ahmad, G.N., Shafiullah, Fatima, H., Abbas, M., Rahman, O., Imdadullah, *et al.* (2022) Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features. *Applied Sciences*, **12**, Article 7449. https://doi.org/10.3390/app12157449

[6] Agrawal, Y., Kumar, M., Ananthakrishnan, S. and Kumarapuram, G. (2022) Evapotranspiration Modeling Using Different Tree Based Ensembled Machine Learning Algorithm. *Water Resources Management*, **36**, 1025-1042. https://doi.org/10.1007/s11269-022-03067-7

[7] Loukika, K.N., Keesara, V.R. and Sridhar, V. (2021) Analysis of Land Use and Land Cover Using Machine Learning Algorithms on Google Earth Engine for Munneru River Basin, India. *Sustainability*, **13**, Article 13758.

https://doi.org/10.3390/su132413758

[8] Yang, R. and Yu, Y. (2021) Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis. *Frontiers in Oncology*, **11**, Article 638182. https://doi.org/10.3389/fonc.2021.638182

[9] Dabija, A., Kluczek, M., Zagajewski, B., Raczko, E., Kycko, M., Al-Sulttani, A.H., *et al.* (2021) Comparison of Support Vector Machines and Random Forests for Corine Land Cover Mapping. *Remote Sensing*, **13**, Article 777. https://doi.org/10.3390/rs13040777

[10] Avci, C., Budak, M., Yağmur, N. and Balçik, F. (2023) Comparison between Random Forest and Support Vector Machine Algorithms for LULC Classification. *International Journal of Engineering and Geosciences*, **8**, 1-10. https://doi.org/10.26833/ijeg.987605

[11] Subbiah, S.S. and Chinnappan, J. (2021) Opportunities and Challenges of Feature Selection Methods for High Dimensional Data: A Review. *Ingénierie des systèmes d information*, **26**, 67-77. https://doi.org/10.18280/isi.260107

[12] Bharadiya, J.P. (2023) A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning. *International Journal of Innovative Science and Research Technology*, **8**, 2028-2032.

[13] Ayon, S.I., Islam, M.M. and Hossain, M.R. (2020) Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. *IETE Journal of Research*, **68**, 2488-2507. https://doi.org/10.1080/03772063.2020.1713916

[14] Khan, A., Qureshi, M., Daniyal, M. and Tawiah, K. (2023) A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction. *Health & Social Care in the Community*, **2023**, Article ID: 1406060. https://doi.org/10.1155/2023/1406060

[15] Bajaj, S. and Behera, L. (2023) Predictive Modeling of Cardiovascular Disease Using Machine Learning Techniques. 2023 *Third International Conference on Secure Cyber Computing and Communication* (*ICSCCC*), Jalandhar, 26-28 May 2023, 518-523. https://doi.org/10.1109/icsccc58608.2023.10176425

[16] Asif, D., Bibi, M., Arif, M.S. and Mukheimer, A. (2023) Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. *Algorithms*, **16**, Article 308. https://doi.org/10.3390/a16060308

[17] Singh, P., Pal, G.K. and Gangwar, S. (2022) Prediction of Cardiovascular Disease Using Feature Selection Techniques. *International Journal of Computer Theory and Engineering*, **14**, 97-103. https://doi.org/10.7763/ijcte.2022.v14.1316

[18] Krittanawong, C., Virk, H.U.H., Bangalore, S., Wang, Z., Johnson, K.W., Pinotti, R., *et al.* (2020) Machine Learning Prediction in Cardiovascular Diseases: A Meta-Analysis. *Scientific Reports*, **10**, Article No. 16057. https://doi.org/10.1038/s41598-020-72685-1

[19] Tsarapatsani, K., Sakellarios, A.I., Pezoulas, V.C., Tsakanikas, V.D., Kleber, M.E., Marz, W., *et al.* (2022) Machine Learning Models for Cardiovascular Disease Events Prediction. 2022 44*th Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (*EMBC*), Glasgow, 11-15 July 2022, 1066-1069. https://doi.org/10.1109/embc48229.2022.9871121

[20] Dalal, S., Goel, P., Onyema, E.M., Alharbi, A., Mahmoud, A., Algarni, M.A., *et al.* (2023) Application of Machine Learning for Cardiovascular Disease Risk Prediction. *Computational Intelligence and Neuroscience*, **2023**, Article ID: 9418666. https://doi.org/10.1155/2023/9418666

[21] Yadav, A.L., Soni, K. and Khare, S. (2023) Heart Diseases Prediction Using Machine

Learning. 2023 14*th International Conference on Computing Communication and Networking Technologies* (*ICCCNT*), Delhi, 6-8 July 2023, 1-7. https://doi.org/10.1109/icccnt56998.2023.10306469

[22] Bhatt, C.M., Patel, P., Ghetia, T. and Mazzeo, P.L. (2023) Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, **16**, Article 88. https://doi.org/10.3390/a16020088

[23] Riyaz, L., Butt, M.A., Zaman, M. and Ayob, O. (2021) Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. In: Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand, S. and Jaiswal, A., Eds., *International Conference on Innovative Computing and Communications*, Springer, 81-94. https://doi.org/10.1007/978-981-16-3071-2_8

[24] Javaid, M., Haleem, A., Singh, R.P. and Suman, R. (2021) Artificial Intelligence Applications for Industry 4.0: A Literature-Based Study. *Journal of Industrial Integration and Management*, **7**, 83-111. https://doi.org/10.1142/s2424862221300040

[25] Touretzky, D., Gardner-McCune, C. and Seehorn, D. (2022) Machine Learning and the Five Big Ideas in Ai. *International Journal of Artificial Intelligence in Education*, **33**, 233-266. https://doi.org/10.1007/s40593-022-00314-1

[26] Sjödin, D., Parida, V., Palmié, M. and Wincent, J. (2021) How AI Capabilities Enable Business Model Innovation: Scaling AI through Co-Evolutionary Processes and Feedback Loops. *Journal of Business Research*, **134**, 574-587. https://doi.org/10.1016/j.jbusres.2021.05.009

[27] Das, A. (2023) Logistic Regression. In: Maggino, F., Ed., *Encyclopedia of Quality of Life and Well-Being Research*, Springer, 3985-3986. https://doi.org/10.1007/978-3-031-17299-1_1689

[28] Cunningham, P. and Delany, S.J. (2021) K-Nearest Neighbour Classifiers—A Tutorial. *ACM Computing Surveys*, **54**, 1-25. https://doi.org/10.1145/3459665