AIMS Medical Science

AIMS Medical Science, 11(2): 58–71.
DOI: 10.3934/medsci.2024005
Received: 27 January 2024
Revised: 11 April 2024
Accepted: 23 April 2024
Published: 07 May 2024

http://www.aimspress.com/journal/medicalScience

Research article

Enhancing cardiovascular disease prediction: A hybrid machine learning approach integrating oversampling and adaptive boosting techniques

Segun Akinola[1], Reddy Leelakrishna[2] and Vijayakumar Varadarajan[3,*]

[1] Johannesburg Business School, University of Johannesburg, ZA-2006, South Africa
[2] Department of Physics, University of Johannesburg, ZA-2006, South Africa
[3] Ajeenkya DY Patil University, Pune, Maharashtra, India

* **Correspondence:** Email: dean.international@adypu.edu.in.

Abstract: This study presents a novel approach to enhance cardiovascular disease prediction using a hybrid machine learning (ML) model. Leveraging on Synthetic Minority oversampling techniques (SMOTE) and adaptive boosting (AdaBoost), we integrate these methods with prominent classifiers, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Extra Tree (ET). Focused on heart rate data as stress level indicators, our objective is to jointly predict cardiovascular disease, thereby addressing the global health challenge of early detection and accurate risk assessments. In response to class imbalance issues in cardiology databases, our hybrid model, which combines SMOTE and AdaBoost, demonstrates promising results. The inclusion of diverse classifiers, such as RF, XGBoost, and ET, enables the model to capture both linear and nonlinear relationships within the heart rate data, significantly enhancing the prediction accuracy. This powerful predictive tool empowers healthcare providers to identify individuals at a high risk for heart disease, thus facilitating timely interventions. This article underscores the pivotal role of ML and hybrid methodologies in advancing health research, particularly in cardiovascular disease prediction. By addressing the class imbalance and incorporating robust algorithms, our research contributes to the ongoing efforts to improve predictive modeling in healthcare. The findings presented here hold significance for medical practitioners and researchers striving for the early detection and prevention of cardiovascular diseases.

Keywords: stress level indicators; class imbalance; early detection; risk assessment; health research

1. Introduction

Application of machine learning (ML) techniques in healthcare has shown a great potential for accurate diagnoses and the prediction of disease outcomes alongside an integration of multiple concepts. ML hybrid approaches have emerged as a promising approach to enhance health research [1]. This article focuses on the use of a specific hybrid methodology: a combination of Synthetic Minority oversampling techniques (SMOTE) and adaptive boosting (Ada-Boost). Moreover, there are comprehensive classifiers, including random forest (RF), Extreme Gradient Boosting (XGBoost), and Extra Tree (ET), to predict cardiovascular disease (i.e., heart disease) using heart rate data linked to various stress levels. Heart disease is the leading cause of death globally, with early prognoses and accurate cardiovascular risk assessments bring critical for timely interventions and effective treatments [2]. ML system types enable the investigation of large amounts of heart rate data and the identification of patterns and factors that may contribute to the prediction of cardiovascular outcomes. The combination of SMOTE and AdaBoost in a machine learning hybrid approach holds a particular promise for cardiovascular disease prediction [3]. SMOTE addresses the issue of class imbalances in cardiology databases, where the number of positive (disease) cases is generally low compared to negative (non-disease) cases. It combines a strong classifier and uses the strength of each classifier to improve the overall prediction accuracy. The different classifiers used in this hybrid approach with different algorithms are known for their capability in handling high-dimensional data with separate classes by finding the best hyperplane [4]; this is an excellent statistical modelling technique in binary distribution problems and provides interpretable results. RF and XGBoost are ensemble algorithms that combine more than one models to form a robust predictive model, and ET algorithms provide interpretations and transformations to capture feature interactions. Combining these classifications in a hybrid system offers several advantages in the prediction of cardiovascular disease [5]. Combining the strengths of each algorithm, the hybrid model can effectively capture both linear and nonlinear relationships in heart rate data, thus resulting in accurate and reliable predictions. This enables healthcare providers to identify individuals with an increased risk of cardiovascular disease, resulting in an active intervention [6]. This article focuses on investigating the possible use of hybrid ML techniques to predict cardiovascular disease caused by various stress levels using heart rate data. We investigate the methodology, implementation, and performance, and evaluate the hybrid approach with SMOTE, Ada-Boost, and classifiers (RF, XGB and ET). In addition, we discuss the importance of feature selection and model interpretation in cardiovascular prognosis. By using hybrid ML techniques, healthcare professionals can unlock valuable insights from heart rate data, thus enabling earlier and more accurate diagnoses and predictions of heart disease. This has the potential to transform the healthcare practice through early intervention, improved patient outcomes, and an improved distribution of healthcare resources. The usage of hybrid ML techniques, especially the combination of SMOTE with Ada-Boost with advanced classifiers, offers a powerful approach to cardiovascular disease prediction using heart rate data. The combination of these methods addresses class imbalance issues, captures complex patterns, and increases the prediction accuracy. By integrating RF, XGBoost, and ET into the hybrid systems, health care professionals can make more rational decisions, thus leading to earlier cardiovascular disease diagnoses and improved management plans. The major contribution of this research work includes the following:

(1) The article contributes to integrating ML and hybrid models by emphasizing the importance of using ML and hybrid methods for health research, especially for cardiovascular disease prediction.

The combination of SMOTE and AdaBoost with accurate classifiers enables healthcare professionals to harness the power of multiple algorithms to improve the prediction accuracy and reliability.

(2) The article highlights the importance of preventing a class imbalance in cardiovascular disease prognosis. Using the SMOTE oversampling approach, hybrid models are able to handle imbalanced data sets, thus improving the detection of cardiovascular events. This contribution increases the reliability of predictions and reduces the death risk due to timely interventions.

(3) The case improves the diagnoses by introducing a variety of algorithms including RF, XGBoost, and ET in the hybridization process. Each algorithm brings unique strengths, such as high-quality data handling, object relations capture, and interpretations. Combining these classifications enables an improved analysis of the heart rate data and enhances the predictive performance of the hybrid models.

(4) The article focuses on the use of cardiac arrhythmia data specific to physiology. Using ML, health care professionals have gained valuable insights from heart rate data and identified individuals at a high risk for heart disease. These contributions enable early intervention and personalized treatment strategies, ultimately improving patient outcomes.

(5) The article offers insights into the methodology and application of the ML hybrid approach. It provides guidance on integrating SMOTE, AdaBoost, and the classifier sets, including practical considerations of the feature selection and model interpretation capabilities. This insight enables researchers and healthcare professionals to effectively use hybrid methods in cardiovascular prognoses and obtain meaningful results.

Finally, the contribution of the article is based on integrating ML with hybrid methods to handle class imbalances using a wider range of classifiers, which focused on controlled heart and heart rate prediction data, and practical insights are provided into their use.

1.1. Related work

Cardiovascular diseases pose a major human health hazard, where the disease death rate keeps increasing daily There are numerous ways to reduce the death rate of heart disease, which include early prediction and detection; researchers across the globe are identifying various novel approaches that will either assist or prevent the death rate. This includes the accurate use of several supervised ML algorithms to understand vascular disease rates using various dataset. The applied classifiers involved artificial neural networks (ANN), RF, and Naïve bayes. The result was computed and showed that ANNs obtained the highest accuracy at 94.1% alongside a sensitivity and specificity of 94.1% [7]. In [8], ML approaches for heart diseases prediction was performed using medical data and past information; several techniques were discussed and compared in the paper, which included five common plans to predict the chance of heart disease within the analyses. The ML used the k-nearest neighbor (KNN). This was performed with python programming within the Jupiter notebook; the result showed that the KNN had an accuracy of both 87% and 79%. In [9], a proactive study using an ML algorithm to predict heart disease was performed with a University of California, Irvine (UCI) data set and the KNN. The study revealed an accuracy of 80%, a KNN of 85%, 73%, and 78% in the accuracies respectively. The model showed that the KNN performed the best based on its highest accuracy within the outcome. The effect of heart disease predictions with hybrid ML techniques was performed in [10], in which the proposed model aimed to significantly find features with the application of ML to enhance the results accuracy prediction of cardiovascular disease. This model used several combinations of classification techniques such as Naïve bayes, Generalized Linear Models (GLM), Deep Learning (DL), RF, and Gradient Boosting Trees (GBT). The accuracy level was at 88.7% for the hybrid RF and linear model. The research developed a model that accurately predicted heart disease to reduce the fatality caused. The method adopted the k mode clustering hung, which had an enhanced classification accuracy. Models such as RF, Multilayer Perceptron (MP), and XGBoost were used. The model used a 70000 real word data set from the kaggle data site; the model was trained, in which the data which was split 80:20 and obtained an accuracy of 86.37%, with a cross validation of 86.56%, an XGBoost of 86.87% with cross validation and 87.02% without cross validation, an RF of 87% with cross validation and 86% without cross validation, and an MP of 87% with cross validation and 86.94% without cross validation, with the highest accuracy of 87.28% [11]. Experimental results showed the accuracy of the extra trees classifier, the logistic model tree classifier, the support vectormachine, and the naive bays classifiers as 90%, 88%, 87%, and 86%, respectively [12]. The performance in the research conducted in [13] was measured in terms of the accuracy, sensitivity, specificity, F1-score, Matthews Correlation Coefficient (MCC) value, and area under the curve (AUC) value. Two features were selected by both feature selection techniques and achieved the highest overall accuracy of 80% for the decision tree classifier.

1.2. Background on machine learning

ML has become central to artificial intelligence, enabling computers to learn and make informed or predictive decisions without explicit programming. Its applications span areas as diverse as healthcare, finance, and image recognition. ML algorithms are designed to analyze complex patterns in data, thus enabling models to accurately classify results. Learning from historical data, these algorithms can develop their knowledge in general and apply it to unseen information. By statistical methods, it also has the ability to make reliable predictions by mathematical modelling and optimization algorithms.

1.3. Synthetic minority oversampling technique (SMOTE)

This is an oversampling method adopted to address the class imbalances in cardiovascular prognostic data sets [14]. It creates an artificial model for a subset (cardiac cases) by identifying the feature space of existing models. The mathematical procedure for SMOTE involves selecting a subclass instance and its k nearest neighbors [15]; then, new instances of products are created on line segments that connect them. This method aims to stabilize the class dissemination to enhance the cardiovascular representation in dataset, thus improving the outcome of the prediction models. For the heart disease prediction, we focus on identifying instances that either indicate heart disease ($y = 1$) or no heart disease ($y = 0$) with a high relevance ($\varphi(y) > tE$). We separately generate synthetic cases for both categories and concatenate them. The undersampling step involves randomly selecting cases from the remaining dataset (excluding instances indicating heart disease and no heart disease) to achieve a class balance. The Algorithm 1 (See Appendix 1).

1.4. Extreme gradient boosting (XGBoost)

This is an advanced cluster learning algorithm known for its high prediction performance [16]. XGBoost combines the strengths of gradient enhancements and regularization techniques to increase the accuracy of heart prediction models. The algorithm uses a gradient-based optimization approach, where each subsequent tree is trained to correct the errors of previous trees. The mathematical model includes minimizing the loss function that measures the difference among predicting cardiovascular outcomes with actual outcomes [17], XGBoost uses a regularization step to control the model complexity and prevent overfitting, thus resulting in an improved prediction accuracy, as calculated by Algorithm 2 (see Appendix 1).

1.5. Extra tree (ET)

This is a ML algorithm used for prediction tasks [18]. It is an ensemble technique that forms and combines their predictions to make accurate predictions. Unlike the traditional RF, ETs introduces additional randomness by randomly selecting features and split points during the construction of each tree. This randomness helps to decrease overfitting while improving the model's robustness. ETs are known for their proficiency in handling datasets. It has been generally applied in several domains, such as healthcare, finance, and natural language processing, and is calculated by Algorithm 3 (see Appendix 1).

1.6. Random forest (RF)

This is a powerful ML technique widely used to predict stress-induced heart disease [19]. It works by constructing various decision trees and combining their predictions into plausible and robust classifications. RF overpowers the limits of singular decision trees by reducing the overfitting while improving the generalization. Using RF selection and bootstrap aggregation, RF captures strong associations between stress-related factors and cardiovascular diseases such as hypertension, cholesterol levels, and lifestyle choices [20]. It makes this an effective tool and can be calculated by Algorithm 4 (see Appendix 1).

1.7. Adaptive boosting (AdaBoost)

AdaBoost is an ensemble learning algorithm which constructs a robust classifier for cardiovascular prediction by combining multiple weak classifiers [21]. The algorithm starts by assigning equal weights to all training samples. It selects the weakest classifier at separate iterations and adjusts the weights of well-classified samples to highlight complex cases. The final predicted model is obtained by weighting the predicted weak distributions by their performance. AdaBoost optimizes weights and merges classifiers more frequently, thus incrementally improving cardiovascular prediction accuracy [22], and can be calculated by Algorithm 5 (see Appendix 1).

2. Materials and methods

This article presents a method for cardiovascular disease prediction using a combination of oversampling techniques (SMOTE) and adaptive boosting (AdaBoost), as well as other ML classifiers such as RF, XGBoost, and ET [23]. The goal is to harness the strengths of these methods to increase the accuracy and robustness of cardiovascular predicting models. The procedure follows the following steps (see Figure 1).

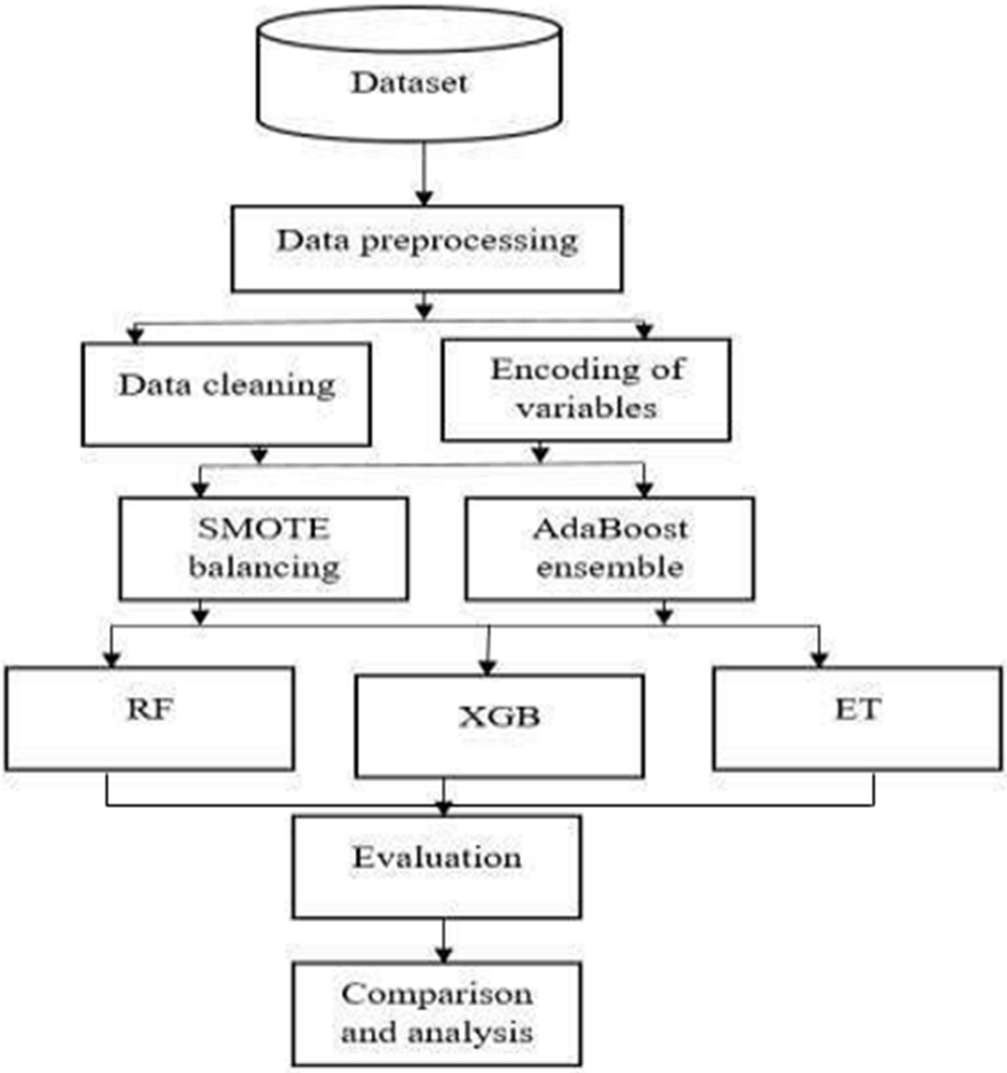


Figure 1. Methodology process.

(1) Data pre-processing: The cardiovascular disease dataset is obtained from reliable sources and undergoes pre-processing stages such as data cleaning, handling of missing values, and the encoding of categorical variables. Feature selection techniques are used to identify more information for prognoses.

(2) Oversampling with SMOTE: Since cardiovascular datasets often suffer from a class imbalance, artificial samples of minority classes are created using the SMOTE method, which synthesizes new

instances by interpolating between existing instances of the same class; moreover, it balances the dataset well [24].

(3) AdaBoost is used as an ensemble method to combine multiple weak classifiers and build complex predictive models. It assigns higher weights to well-classified data, thus allowing subsequent classifiers to focus on that data and improve the overall forecast performance [25].

(4) Applications of classifiers: Various classifiers such as RF, XGBoost, and ET are used for cardiovascular prediction. Each classifier is trained on a reprocessed data set and oversampled data obtained from the SMOTE method. The parameters of the classifiers are optimized using techniques such as cross-validation and grid searches to improve their performance.

(5) Evaluation criteria: The performance of the prediction model is evaluated using numerous evaluation metrics such as accuracy, precision, re-call, F1 score, and so on.

(6) Comparison and analysis: The results of each classification were compared and analyzed to determine the most effective model for cardiovascular predictions. The strengths and weaknesses of each approach are explored, and insights are drawn about its suitability for a given data structure [26].

2.1. Dataset

The experimental dataset used here was generated from the kaggle website (Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction, 2023), which includes various attributes taken from signals measured using electrocardiograms (ECG) recorded for different persons having different heart rates at the period the measurement was taken. These various features contribute to the heart rate at the given instant of time for the individual include the following: a specific Unique ID for each patient; the state of the patient at the time the data was documented; the ID of the entire dataset; the higuci fractal dimension of the heart rates; the Poincaré plot standard deviation along the line of identity; the Poincare plot standard deviation perpendicular to the line of identity; and the sample entropy, which measures the consistency and complexity of a time series. This dataset had 41034 records and was used to predict heart disease based on stress, which could ultimately lead to death.

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

$$\frac{\text{Sensitivity}}{\text{recall}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

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

FI = 2 × Precision × Sensitivity, Precision + Sensitivity.

FI scores were used, where TP and TN represented a true positive and true negative in Equations 1, 2 and 3, respectively; this was used to represent a patient with heart stress and no stress, in which the numbers of stress were positive and no stress were negative. This was correctly classified and either the stress or no stress was predicted.

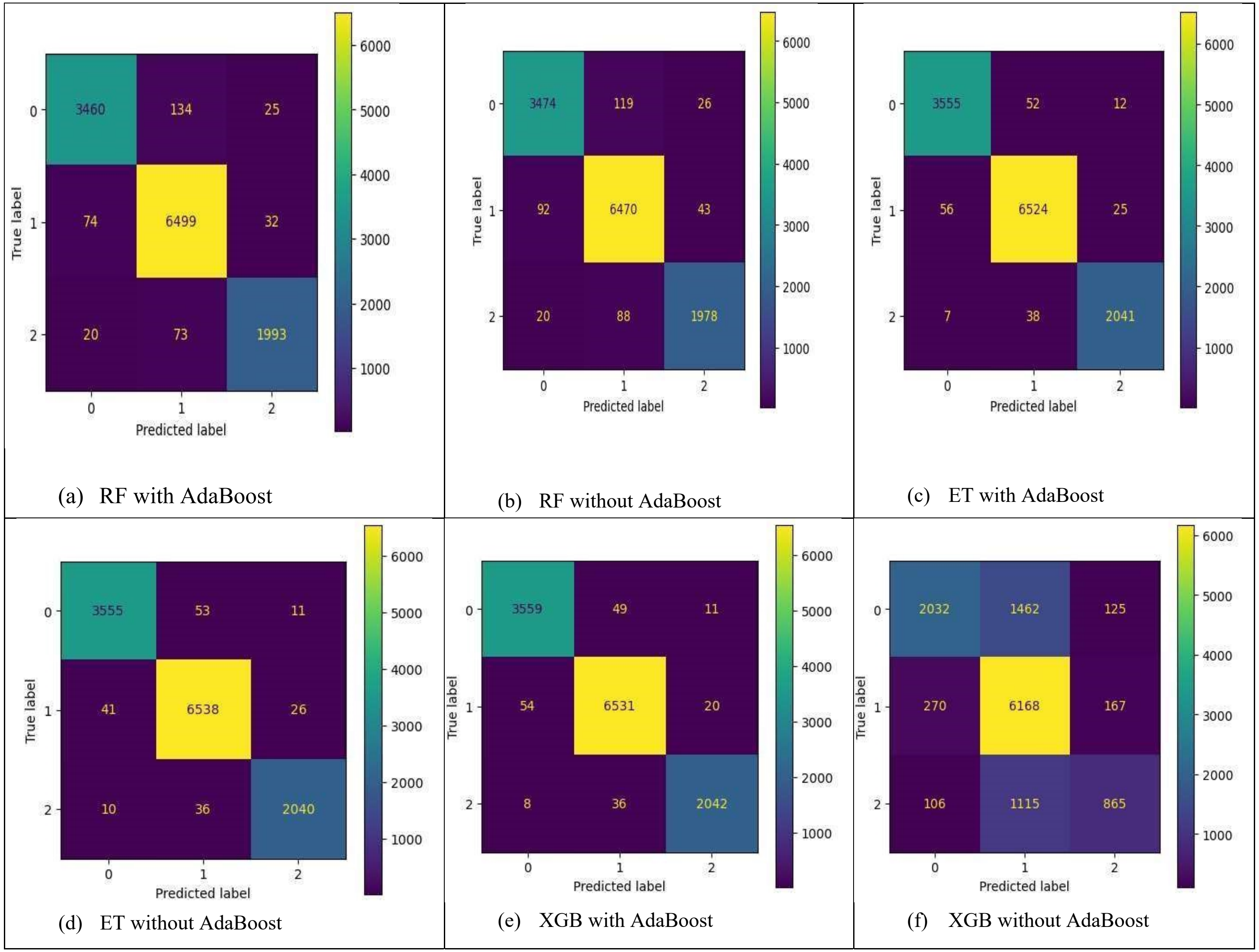
2.2. Experimental set up

We obtained a reliable data set with a large number of documented examples of cardiovascular disease, thus ensuring that the data were relevant to various demographics and clinical factors related to cardiovascular health. In addition, we performed important preliminary steps on the dataset, including handling missing values, data cleaning, and normalization. The data set was divided into smaller training testing units, typically using 70:30 or 80:20 ratios. The experiment was conducted using a computer with the following specifications: Intel (R) Core (TM) i72860QM CPU processor @ 2.50 GHz, RAM 16.0 GB, and a 64-bit operating system, x64-based. Python pro-gramming use with the hybrid ML implementation.

This paper used a large data set that was publicly available, which could be found at (kaggle.com). Utilizing freely available public datasets enhances research accessibility, facilitates reproducibility, and fosters collaboration. Access to diverse, pre-existing data accelerates analyses, reduces costs, and enables comparisons across studies, thus enriching scientific inquiry and promoting innovation.

3. Results

The experimental result revealed an RF accuracy of 96.84, a recall (RC) of 95.99%, a precision of 96.87%, a Matthews correlation coefficient of 0.94, and the Confusion Matrix (CM) was computed to assert the algorithm (Figure 2a). The AdaBoost-SMOTE with RF had an improved accuracy of 97.09, a recall of 95.60%, a precision of 97.35%, and a Matthews correlation coefficient of 0.95. This show improvement with the AdaBoost-SMOTE, as seen within the CM (see Figure 2b). The ET result without AdaBoost showed an RF accuracy of 98.56, a recall of 98.23%, a precision of 98.58%, a Matthews correlation coefficient of 0.97, and the CM was computed to assert the algorithm (see Figure 2c). The AdaBoost-SMOTE with ET had an improved accuracy of 98.45, a recall of 98.23%, a precision of 98.25%, and a Matthews correlation coefficient of 0.97, which is shown within the CM (see Figure 2d). The XGBoost without AdaBoost demonstrated an RF accuracy of 73.63, a recall of 56.14%, a of 84.38%, a Mattews correlation coefficient of 0.54, and the CM was computed to assert the algorithm (see Figure 2e). The AdaBoost-SMOTE with XGB had an improved accuracy of 98.55, a recall of 98.34%, a precision of 98.28%, and a Matthews correlation coefficient of 0.97, which is shown with the CM (see Figure 2f). Table 1 displays the computed summary without AdaBoost and Table 2 displays the computed summary with AdaBoost.



(a) RF with AdaBoost

(b) RF without AdaBoost

(c) ET with AdaBoost

(d) ET without AdaBoost

(e) XGB with AdaBoost

(f) XGB without AdaBoost

Figure 2. Experimental results.

Table 1. Results without ADB-SMOTE techniques.

Model	MC	PR	AC	RC
RF	0.94	96.87	96.84	95.99
ET	0.97	98.58	98.56	98.23
XGB	74.62	84.38	73.63	56.14

Note: ADB-SMOTE: Adaptive boosting-synthetic minority oversampling technique; MC: Matthews correlation coefficient; PR: Precision; AC: Accuracy; RC: Recall; RF: Random forest; ET: Extra tree; XGB: Extreme gradient boosting.

Table 2. Results with ADB-SMOTE techniques.

Model	AC	RC	PR	MC
RF	97.09	95.60	97.35	0.95
ET	98.45	98.23	98.25	0.97
XGB	98.55	98.34	98.28	0.97

Note: ADB-SMOTE: Adaptive boosting-synthetic minority oversampling technique; AC: Accuracy; RC: Recall; PR: Precision; MC: Matthews correlation coefficient; RF: Random forest; ET: Extra tree; XGB: Extreme gradient boosting.

4. Comparison

The comparison analysis was performed on the proposed model with the existing literature, where the results of the existing related cases in the literatures was examined. This shows that the proposed model performs well with AdaBoost in terms of the accuracy; moreover, there is a larger significance with the XGBoost compared to the others. This model showcases a potential to predict health-related cases, as shown in Table 3. The selected algorithms—SMOTE for addressing class imbalance, AdaBoost for ensemble learning, RF for robustness, XGBoost for scalability, and ET for added randomness—suit cardiovascular disease prediction due to their efficiency, interpretability, and ability to handle complex, high-dimensional data. Compared to deep learning models, they offer an improved interpretability and require less data for effective training.

Table 3. Comparison.

AC%	Model	Ref
94.1	ANN	[7]
87	KNN	[8]
78	KNN	[9]
88.7	RF	[10]
87.2	RF	[11]
86	ET	[12]

Continued on next page

AC%	Model	Ref
86	LR	[13]
97.09	RF-AdaBoost	Proposed model
98.45	ET-AdaBoost	Proposed model
98.55	XGB-AdaBoost	Proposed model

Note: AC: Accuracy; ANN: Artificial neural networks; KNN: K-nearest neighbor; RF: Random forest; ET: Extra tree; LR: Logistic regression; RF-AdaBoost: Random forest with AdaBoost; ET-AdaBoost: Extra trees with AdaBoost; XGB-AdaBoost: XGBoost with AdaBoost.

4.1. Experimental validation

This experiment was validated using another data set, namely the HIV prediction model (https://www.kaggle.com/datasets/ishigamisenku10/hiv-prediction, 2023) [27]. The model performance was tested on HIV data using the same ML methods including SMOTE, AdaBoost, RF, XGBoost, and ET. This validation helps to check the generalizability of the retrospective model and the reliable prediction of HIV-related outcomes, this supporting its potential use in health care (see Appendix 2 and Table 4).

Table 4. Validation table with AdaBoost-SMOTE.

Model	AC	RC	PR	MC
RF	86.71	78.26	84.37	0.69
ET	86.85	81.15	84.84	0.72
XGB	88.0	81.15	87.5	0.74

Note: AdaBoost-SMOTE: Adaptive boosting-synthetic minority oversampling technique; AC: Accuracy; RC: Recall; PR: Precision; MC: Matthews correlation coefficient; RF: Random forest; ET: Extra tree; XGB: Extreme gradient boosting.

5. Conclusions

In conclusion, this study demonstrates the effectiveness of ML hybrid approaches to use on cardiovascular disease heart rate data. The combination of SMOTE and AdaBoost solves class imbalance problems and improves the prediction accuracy. The combination of the RF, XGBoost, and ET classifiers allowed one to capture both linear and nonlinear relationships, thus increasing the reliability of the predictions. The findings highlight the importance of ML and hybrid methods in health research, especially in the prediction of cardiovascular disease.

5.1. Future work

Identifying additional variables related to heart health, such as blood pressure, cholesterol levels, and lifestyle factors, can increase the predictive power of the model. If these variables are included, they can lead to accurate risk assessments and personalized disease prevention recommendations.

Expanding research to examine detailed heart rate data could provide insight into disease progression and enable the development of dynamic predictive models. Measuring changes in heart rates over time could enhance the model's ability to detect warning signs of heart failure early and help improve timely interventions. Continuing to advance these research areas, we can increase the accuracy and applicability of ML models to predict cardiovascular disease and ultimately drive patient care to improve and reduce the global burden of cardiovascular disease morbidity and mortality.

Author contributions

Segun Akinola: Conceptualization. Reddy Leelakrishna: Provided the data support. Vijayakumar Varadarajan: Editing and correction.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Data availability

Supplementary materials: All data used are Available: Kaggle website (https://www.kaggle.com).

Conflict of interest

The authors declare no conflicts of interest.

References

1. Sathya D, Sudha V, Jagadeesan D (2020) Application of machine learning techniques in healthcare. In: *Handbook of Research on Applications and Implementations of Machine Learning Techniques*, Pennsylvania: IGI Global, 289–304.
2. Vogel B, Acevedo M, Appelman Y, et al. (2021) The Lancet women and cardiovascular disease commission: reducing the global burden by 2030. *Lancet* 397: 2385–2438. https://doi.org/10.1016/S0140-6736(21)00684-X
3. Benhar H, Idri A, Fernández-Alemán JL (2020) Data preprocessing for heart disease classification: A systematic literature review. *Comput Methods Programs Biomed* 195: 105635. https://doi.org/10.1016/j.cmpb.2020.105635
4. Aljarah I, Al-Zoubi AM, Faris H, et al. (2018) Simultaneous feature selection and support vector machine optimization using the grasshopper optimization algorithm. *Cogn Comput* 10: 478–495. https://doi.org/10.1007/s12559-017-9542-9
5. Woźniak M, Grana M, Corchado E (2014) A survey of multiple classifier systems as hybrid systems. *Inform Fusion* 16: 3–17. https://doi.org/10.1016/j.inffus.2013.04.006
6. Burke LE, Ma J, Azar KMJ, et al. (2015) Current science on consumer use of mobile health for cardiovascular disease prevention: A scientific statement from the American Heart Association. *Circulation* 132: 1157–1213. https://doi.org/10.1161/CIR.0000000000000232

7. Muhammed SM, Abdul-Majeed G, Mahmoud MS (2023) Prediction of heart diseases by using supervised machine learning algorithms. *Wasit J Pure Sci* 2: 231–243.
8. Singh A, Kumar R. Heart disease prediction using machine learning algorithms. In 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452–457, https://doi.org/10.1109/ICE348803.2020.9122958
9. Memon B, Ghulamani S (2022) A relative study of different machine learning classification algorithms to forecast the heart disease. *J Inf Sy Digit Tec* 4: 11–27. https://doi.org/10.31436/jisdt.v4i1.305
10. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7: 81542–81554, https://doi.org/10.1109/ACCESS.2019.2923707
11. Shafique R, Mehmood A, Ullah S, et al. (2019) Cardiovascular disease prediction system using extra trees classifier. https://doi.org/10.21203/rs.2.14454/v1 (unpublished work)
12. Al Mehedi Hasan M, Shin J, Das U, et al. Identifying prognostic features for predicting heart failure by using machine learning algorithm. In Proceedings of the 2021 11th International Conference on Biomedical Engineering and Technology, Tokyo, Japan, 2021, pp. 40–46, https://doi.org/10.1145/3460238.3460245
13. Goldenberg SL, Nir G, Salcudean SE (2019) A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol* 16: 391–403. https://doi.org/10.1038/s41585-019-0193-3
14. Shuja M, Mittal S, Zaman M (2018) Decision support predictive model for prognosis of diabetes using SMOTE and decision tree. *Int J Appl Eng Res* 13: 9277–9282.
15. Zheng Z, Cai Y, Li Y (2015) Oversampling method for imbalanced classification. *Comput Informa* 34: 1017–1037.
16. Ekanayake IU, Meddage DPP, Rathnayake U (2022) A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Stud Constr Mat* 16: e01059. https://doi.org/10.1016/j.cscm.2022.e01059
17. Vrieze SI (2012) Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol Methods* 17: 228–243. https://doi.org/10.1037/a0027127
18. Wang Z, Mu L, Miao H, et al. (2023) An innovative application of machine learning in prediction of the syngas properties of biomass chemical looping gasification based on extra trees regression algorithm. *Energy* 275: 127438. https://doi.org/10.1016/j.energy.2023.127438
19. Ricciardi C, Cantoni V, Improta G, et al. (2020) Application of data mining in a cohort of Italian subjects undergoing myocardial perfusion imaging at an academic medical center. *Comput Meth Prog Bio* 189: 105343. https://doi.org/10.1016/j.cmpb.2020.105343
20. Cox R (2015) Hegemonic masculinity and health outcomes in men: A mediational study on the influence of masculinity on diet [Dissertation]. The University of Memphis.
21. Rani P, Kumar R, Ahmed NMOS, et al. (2021) A decision support system for heart disease prediction based upon machine learning. *J Reliable Intell Environ* 7: 263–275. https://doi.org/10.1007/s40860-021-00133-6

22. Adeboye NO, Abimbola OV (2020) An overview of cardiovascular disease infection: A comparative analysis of boosting algorithms and some single based classifiers. *Stat J IAOS* 36: 1189–1198. https://doi.org/10.3233/SJI-190609
23. Zhang W, Li H, Hun L, et al. (2022) Slope stability prediction using ensemble learning techniques: A case study in Yunyang County, Chongqing, China. *J Rock Mech Geotech* 14: 1089–1099. https://doi.org/10.1016/j.jrmge.2021.12.011
24. Chawla NV, Bowyer KW, Hall LO, et al. (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 16: 321–357. https://doi.org/10.1613/jair.953
25. Breiman L (2001) Random forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324
26. Kaggle, Heart Failure Prediction Dataset. San Francisco Kaggle, 2021. Available from: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. Accessed June 08, 2023.
27. Kaggle, HIV AIDS Dataset. San Francisco Kaggle. Available from: https://www.kaggle.com/datasets/imdevskp/hiv-aids-dataset. Accessed June 08, 2023.