

An Integrated Machine Learning and Deep Learning Framework for Predicting Cardiovascular Disorder

Nirmala M^{1*}, and M. Bhargavi Krishna²

^{1,2}Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Karkambadi, Tirupati, India.

Abstract. One of the biggest and most prominent cardiac conditions that affects humans of any age is a heart attack. Physicians must be accurate in their findings since they work with the lives of people, which is priceless. If people consume their drugs and therapy diligently according to the schedule, early identification of cardiovascular illness may extend many lives. Earlier approaches for predicting cardiovascular illnesses were helpful in decision-making regarding the adjustments that should have taken place in people at greatest risk, thereby lowering their risks. Machine learning (ML) algorithms are necessary to make accurate choices in the forecasting of cardiac problems because the healthcare sector has a large amount of clinical information. In order to classify coronary illness, this research compares machine learning methods such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Artificial Neural Network (ANN). The algorithms with the most significant accuracy in forecasting were LR, RF, SVM, XGBoost, and ANN (80.33%, 81.97%, 80.33%, 78.69%). These findings suggested that RF and SVM are the best techniques for CVD forecasting and categorization. The ML methods Random Forest and SVM is utilized in the suggested approach since it has been shown through a contrast investigation to possess the most highly precise and dependable technique.

1 Introduction

Coronary artery disease and stroke are among the most prevalent forms of cardiovascular diseases (CVDs), which are the largest global cause of mortality. According to the World Health Organization (WHO), cigarette smoking, heavy drinking, inadequate nutrition, and lack of exercise are significant psychological hazards. All of these variables lead to supplementary hazards, such as being overweight and having high blood sugar levels, that are controlled with drugs and dietary modifications to avoid dying young. The primary cause of cardiovascular diseases (CVDs), which account for almost 32% of all fatalities globally in 2022 (an expected 19.8 million dead), is gradually increasing the global mortality numbers annually [1-2].

Cardiovascular disease (CVD) is a serious illness that is becoming more common in all countries. Although it can support physicians' attempts to create a more favorable environment for treatment and diagnosis for patients, the ability of predictive machine learning algorithms to shape doctors' perceptions is crucial to all parties involved in the medical community [3].

Providing treatment is an essential part of life for people. A variety of illnesses influencing the cardiac system and artery walls fall under the general category of heart attack or stroke. Earlier approaches for predicting cardiovascular illnesses were helpful in decision-making regarding the adjustments that should have taken place in people at greatest risk, thereby lowering their risks. ML models are necessary to make

accurate choices in the forecasting of cardiac problems because the healthcare sector has a large amount of clinical information [4-6].

The World Health Organization (WHO) has reported that heart disorders continue to be among the primary causes of death globally, taking the lives of about 17.9 million people each year. The variety of risk variables and the expanding population make early identification more difficult. Data-driven forecasting algorithms made possible via the latest advances in machine learning (ML) has made a significant influence to medical science. The UCI Heart Disease information set, which included 14 medical variables, was employed in this study to create a predictive algorithm for cardiac arrest. Random Forest showed better accuracy and processing efficiency than the other techniques, suggesting that it might be used as a medical decision-support tool [7]. A large-scale real-world dataset with 70,000 records that was taken from Kaggle and split 80:20 across both test and training sets was used for an investigation by Bhatt et al. [8]. Decision Tree, XGBoost, Random Forest, and Multilayer Perceptron are among the machine learning techniques that have been employed and contrasted. Across those algorithms, the Multilayer Perceptron achieved validation across tests performed the best in classification, exhibiting an efficiency of 87.28%. The assessed algorithms' circumference of the curve (AUC) results, which ranged from 0.94 to 0.95, further indicated their high accuracy in forecasting.

* Corresponding author: nirmalapcr1993@gmail.com

1.1 Related work

Jindal et al. [9] focused on identifying which patients, based on various medical characteristics, are more likely to have cardiovascular disease. Leveraging the individual's medical record, the researchers developed a cardiovascular disease prediction algorithm to determine whether the patient will likely receive a diagnosis of coronary artery disease. To forecast and categorize coronary artery disease patients, they employed various machine learning methods, including logistic regression and KNN. A very useful method was employed to control the exact algorithm might be applied to increase the accuracy of cardiac arrest prediction for each patient. The significance of the suggested approach was quite pleasing; utilizing KNN and Logistic Regression, the approach was enabled to predict signs of cardiovascular illness in a specific person with excellent precision when compared to the classifiers that were earlier employed, like naïve bayes, etc. By applying the provided model to determine the likelihood that the classifier would effectively and precisely detect a cardiac illness, a considerable amount of pressure has been relieved. The Given cardiovascular illness prediction method lowers costs and improves medical treatment. This initiative provides important information that can aid in the prediction of cardiovascular illness cases. Machine learning (ML) is useful for deciding predictions and choices based on the huge amount of information generated by the medical sector. In order to improve the performance of heart attacks prediction, Mohan et al. [10] suggested a unique approach that uses machine learning approaches to identify important variables. Various well-known classification methods and various feature combinations are used for implementing a prediction approach. The hybrid random forest having a linear model (HRFLM) prediction approach to cardiovascular illness has improved efficiency with a precision rate of 88.7%.

The approach based on supervised training methods such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm was provided by Shah et al [11]. along with several characteristics associated with cardiovascular illness. It makes use of the current information of cardiovascular disorder cases from the Cleveland database of the UCI repository. There are 76 characteristics and 303 occurrences in the collection. Only 14 of these 76 qualities are taken into account for investigation, which is crucial to supporting the effectiveness of various algorithms. The purpose of this study is to estimate the likelihood that patients will acquire coronary artery disease. The findings show that K-nearest neighbor yields the most precise value.

Risk assessment, as well as earlier illness identification, are greatly aided by machine learning (ML) techniques. Traditional machine learning methods like Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and Naïve Bayes (NB) have been extensively used for classification challenges in cardiac forecast investigations. Nevertheless, these methods frequently encounter difficulties in efficiently utilizing huge-scale, high-dimensional medical data produced by modern

medical facilities. Moreover, problems with feature complication information variability and approach simplification create it hard to integrate machine learning systems into real medicinal measures, which can irregularly outcome in insufficient prediction precision [12-13].

In order to ascertain the most effective method for predicting illnesses, Bilgaiyan et al. gathered information across four medical centers in four different regions. Additionally, the efficacy of several data mining techniques, such as K-nearest neighbors, Random Forest, Multi-layer Perceptron, and Logistic Regression, for predicting cardiac illness was contrasted. Following an examination of the data mining techniques, the forecast efficacy for each strategy used is presented. The results showed that it is possible to anticipate cardiovascular issues considerably over 97% efficiency [14].

Boukhatem et al. introduced various types of machine learning techniques for predicting cardiovascular illness by utilizing medical information on key clinical variables. Four classification techniques—MLP, SVM, RF, and NB—were used in the study to construct prediction algorithms. The algorithms had been developed once the data was preprocessed and features are chosen. F1-score, recall, accuracy, and precision have been utilized to assess the algorithms. The SVM approach had the highest precision of 91.67% [15].

2 Experimentation

Fig. 1. depicts the heart disease prediction flowchart. The UCI Heart Disease dataset is the starting point for the suggested cardiovascular illness forecasting algorithm. It is preprocessed using duplication elimination, feature scaling, and an 80:20 train-test split. Following preprocessing, a deep learning model (Artificial Neural Network) and four machine learning models (Logistic Regression, Support Vector Machine, Random Forest, and XGBoost) are trained concurrently. The algorithms produce predictions on the test dataset after learning patterns from the training dataset. Metrics including accuracy, precision, recall, F1-score, and ROC-AUC are used to assess their performances. Lastly, the algorithms are compared using confusion matrices and ROC curves to determine which method performs most effectively in predicting cardiovascular illness.

The necessary collection of detailed specifications for each attribute is displayed in Tables 1 and 2. 303 cases via the database were taken into consideration for this investigation. To maintain data validity, one duplication was found and eliminated throughout the data extraction process. Lastly, there were 302 cases for the prediction of heart disease, 164 of which were confirmed, and 138 of which had no symptoms. The features of outcomes from lab tests, clinical assessments, and demographic information are all included in the investigation of heart disease. The root of the problem is given as a kind of binary task from the perspective of machine learning (the target value indicates whether heart illness is present (1) or absent (0)). Create the

feature set using both qualitative and quantitative factors. Appropriate preprocessing steps, such as encoding and normalization, are required before the algorithms are trained. In order to compare efficacy and forecast heart failure, supervised learning approaches are evaluated using datasets gathered from multiple sources [17]. Fig. 2 depicts the workflow of the proposed heart disease prediction model.



Fig. 1. Flow chart of heart disease prediction.

Table 1. Dataset Characteristics for Heart Disease Prediction

Dataset Characteristic	Details
Sample Size	303 patient cases
Feature Count	14 clinical parameters
Duplicates	1
Missing	0
No disease (0)	138
Disease (1)	164
Output Class	1 = heart disease present 0 = heart disease absent
Data Source	UCI Machine Learning Repository (Cleveland subset)
Usage Rights	Open-source (research purposes)

Table 2. Clinical Feature Description of the Heart Disease Dataset

Feature Name	Clinical Interpretation
Age	Patient's age in years
Sex	Gender of the patient (Male = 1, Female = 0)
Chest Pain Type	Category of chest pain symptoms
trestbtps	Resting systolic blood pressure (mm/Hg)
chol	Serum cholesterol concentration (mg/dL)
fbs	Fasting blood sugar status
restecg	Resting electrocardiogram (ECG) findings
thalach	Peak heart rate attained during exercise
exang	Presence of exercise-induced angina (1 = Yes, 0 = No)
oldpeak	ST-segment depression relative to rest
slope	Slope of the ST segment during peak exercise

ca	Number of major coronary vessels identified via fluoroscopy
thal	Thalassemia status (haemoglobin disorder indicator)
Target	Diagnostic outcome (0 = Absence, 1 = Presence of heart disease)

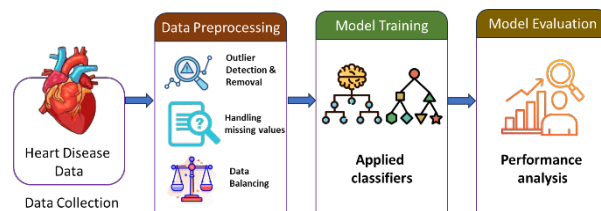


Fig. 2. Workflow of the Proposed Heart Disease Prediction Model

2.1 Novelty

This study is novel because it compares several machine learning and deep learning models within a single, cohesive framework to predict cardiovascular illness. This study compares Logistic Regression, SVM, Random Forest, XGBoost, and ANN using the same preliminary processing and validation approach, unlike conventional research that uses a single method. On an insignificant number of medical datasets, an individual contrast of neural networks with ensemble boosting (XGBoost) yields insightful performance information. Appropriate analysis of models is ensured by accurate assessment with ROC-AUC, F1-score, Accuracy, Precision, and Recall. For an accurate health diagnosis, our hierarchical ML–DL comparison technique improves the choice of models.

3 Results and Discussion

The algorithm is run on Google Colaboratory to calculate the outcomes. In essence, Google Colab enables users to run Python code within a web browser. The first step in the suggested method is to assess the outcomes for each classifier separately. The top-performing classifiers are chosen and evaluated on benchmark datasets related to cardiovascular illness. Cardiovascular illness samples are collected through the machine learning UCI data library, which is accessible online and has been authorized by several investigators. Certain characteristics in the dataset will be utilized to identify whether or not the patient has cardiac problems. Healthier and ill are examples of categorical variables that make up the label attribute, which is essentially the resulting feature. One has been substituted for people in good health values, and zero for the diseased values [18].

The suggested structure for comparison with each classifier for the Cleveland dataset is shown graphically in Fig. 3. In comparison to machine learning and deep learning, the accuracy, precision, recall, F1 score, and AUC for each one and ensemble classifiers are examined. The suggested random forest from machine learning has done noticeably better in terms of Accuracy, Precision, Recall, F1 Score, and AUC, as can be seen from Table 3. Among machine learning

algorithms and deep learning, the suggested RF and SVM from the machine learning system consistently performed well across all models, suggesting that it may be used to predict cardiovascular illness on any dataset. RF and SVM demonstrated an accuracy of above 0.81 across all models, indicating an efficiency level suitable for clinical systems that help make decisions. The suggested machine learning approach's better performance suggests that it might successfully forecast cardiac illnesses [19-20].

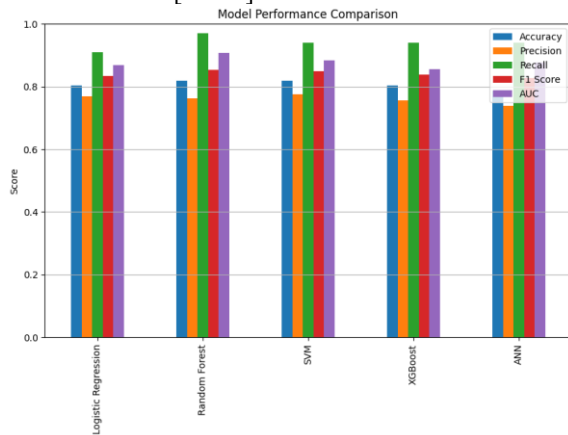


Fig. 3. Algorithm's performance

Table 3. Model performance curve analysis results

Models	Accur acy	Precisi on	Reca ll	F1 Scor e	AU C
Logistic Regress ion	0.8033	0.7692	0.9091	0.8333	0.8690
Rando m Forest	0.8197	0.7619	0.9697	0.8533	0.9080
SVM	0.8197	0.7750	0.9394	0.8493	0.8831
XGBoo st	0.8033	0.7561	0.9394	0.8378	0.8561
ANN	0.7869	0.7381	0.9394	0.8267	0.8766

3.1 ROC curve

Additionally, the computer models' predictive accuracy has been demonstrated using the recursive Operating Characteristic (ROC) curve. The trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (1-specificity) over various threshold values is depicted by the ROC curve. This visual aid provides a thorough assessment of a classifier's discriminative power. The regression coefficient curves associated with the DL and ML classification algorithms are shown in Figure 4. The random forest classifier achieved the highest discriminative efficiency, with an AUC of 0.908, indicating stronger class separation across all

techniques, according to the ROC-AUC evaluation. The Artificial Neural Network (ANN) and Logistic Regression generated comparable AUC values of 0.877 and 0.869, accordingly, whereas the Support Vector Machine (SVM) came in close with an AUC of 0.883. Although XGBoost's AUC was slightly diminished at 0.856, it was substantially predictive beyond the baseline level. Random Forest has the best sensitivity-specificity balance over a range of sensitivities because AUC scores nearer 1 indicate superior trades among the rate of true positives and the rate of false positives [21-22].

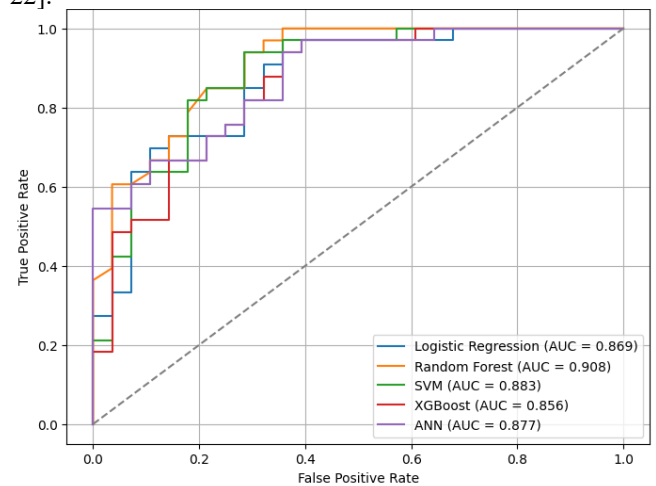


Fig. 4. Recursive operating characteristic curve (ROC) for various classifiers in machine learning and deep learning

3.2 Confusion Matrix

The confusion matrix, which links the real goals for the response factor, Cardiovascular victims, towards the projected outcomes response from the ML and DL model, is employed to assess the effectiveness of the classification algorithm performed. As anticipated, the RF performed superior across all assessment measures for the confusion matrix. By displaying the distribution of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each design, the confusion matrix provides an extensive review of classification quality. Despite 9 false positives and 3 false negatives, the algorithm for logistic regression accurately detected 30 positive instances (TP) and 19 instances of error (TN), showing considerable misclassification in the negative class. RF confirmed great sensitivity and enhanced identification of heart disease patients, attaining 32 TP and 18 TN with only 1 false negative. Further, 31 TP, 19 TN, 2 FN, and 9 FP, the SVM approach revealed reliable efficacy, representing stable classification capability. In comparison to SVM, XGBoost maintained competitiveness but had somewhat poorer specificity, producing 31 TP, 18 TN, 2 FN, and 10 FP. The ANN algorithm had decent sensitivity but comparatively greater false positives, detecting 31 TP and 17 TN with 2 FN and 11 FP. Since reducing the number of undetected cardiovascular illness patients has become medically significant, Random Forest performs most effectively because it has the minimum false negative

rate, that plays an essential role when performing medical early disease identification [23-24].

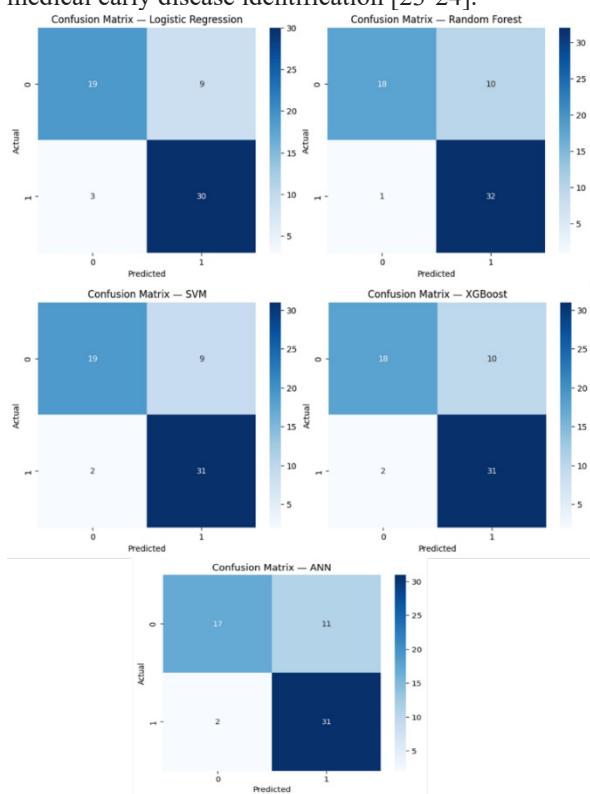


Fig. 5. Confusion matrix of proposed models

4 CONCLUSION

Cardiovascular conditions have been recognized as a major concern while analyzing clinical information. Considering it may support physicians' attempts to create a better environment for patient diagnosis and treatment, the ability of predictive machine learning algorithms to shape physicians' perceptions is crucial for all parties involved in the medical field. The effectiveness of predictive machine learning algorithms for CVD patients was examined in the present investigation. One of the main causes of death globally is cardiovascular disease.

This present research suggests a method for diagnosis that incorporates many pre-trained algorithms, such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Artificial Neural Network (ANN), to solve the problem of timely and efficient cardiac illness identification. Numerous trials on the Cleveland subset of the UCI Machine Learning Repository dataset were used to assess the suggested technique. Outcomes from experiments indicate that the framework performs well in prediction and has little overfitting, achieving high classification accuracy on both training and testing datasets. Across all the algorithms that were assessed, the Random Forest classifier had the best class separability, achieving the highest discriminative performance with an AUC of 0.908, according to the ROC-AUC analysis. The algorithm's significant level of AUC indicates its remarkable capacity to discriminate between different people with cardiovascular illness and those who are not, over a range of decision criteria. Random Forest also showed increased sensitivity, which ensured better heart failure identification while preserving dependable overall classification performance. These results

confirm that the suggested multi-model diagnostic approach for cardiovascular illness forecasting is reliable and efficient.

References

1. N. Chandrasekhar, S. Peddakrishna, Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes* 11, 1210 (2023).
2. J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, G. Wang, A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Med. Eng. Phys.* 105, 103825 (2022).
3. R. Katarya, S. K. Meena, Machine learning techniques for heart disease prediction: A comparative study and analysis. *Health Technol.* 11, 87–97 (2021).
4. A. Singh, R. Kumar, Heart disease prediction using machine learning algorithms. *Proc. Int. Conf. Electr. Electron. Eng. (ICE3)*, 452–457 (2020).
5. V. Sharma, S. Yadav, M. Gupta, Heart disease prediction using machine learning techniques. *Proc. 2nd Int. Conf. Adv. Comput. Commun. Control Netw. (ICACCCN)*, 177–181 (2020).
6. P. Rani, R. Kumar, N. M. S. Ahmed, A. Jain, A decision support system for heart disease prediction based upon machine learning. *J. Reliable Intell. Environ.* 7, 263–275 (2021).
7. L. Riyaz, M. A. Butt, M. Zaman, O. Ayob, Heart disease prediction using machine learning techniques: A quantitative review. *Proc. Int. Conf. Innov. Comput. Commun. (ICICC)*, 81–94 (2021).
8. C. M. Bhatt, P. Patel, T. Ghetia, P. L. Mazzeo, Effective heart disease prediction using machine learning techniques. *Algorithms* 16, 88 (2023).
9. H. Jindal, S. Agrawal, R. Khera, R. Jain, P. Nagrath, Heart disease prediction using machine learning algorithms. *IOP Conf. Ser.: Mater. Sci. Eng.* 1022, 012072 (2021).
10. S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7, 81542–81554 (2019).
11. D. Shah, S. Patel, S. K. Bharti, Heart disease prediction using machine learning techniques. *SN Comput. Sci.* 1, 345 (2020).
12. C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, A. Li, F. Wang, A comprehensive review of deep learning-based models for heart disease prediction. *Artif. Intell. Rev.* 57, 263 (2024).
13. W. A. W. A. Bakar, N. L. N. B. Josdi, M. B. Man, M. A. B. Zuhairi, A review: Heart disease prediction in machine learning and deep learning. *Proc. IEEE Int. Colloq. Signal Process. Appl. (CSPA)*, 150–155 (2023).
14. A. A. Almazroi, E. A. Aldahri, S. Bashir, S. Ashfaq, A clinical decision support system for heart

- disease prediction using deep learning. *IEEE Access* 11, 61646–61659 (2023).
15. M. S. Al Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, A. Shaikh, A robust heart disease prediction system using hybrid deep neural networks. *IEEE Access* 11, 121574–121591 (2023).
 16. S. Bilgaiyan, T. I. Ayon, A. A. Khan, F. T. Johora, M. Parvin, M. J. Alam, Heart disease prediction using machine learning. *Proc. Int. Conf. Comput. Commun. Inform. (ICCCI)*, 1–6 (2023).
 17. C. Boukhatem, H. Y. Youssef, A. B. Nassif, Heart disease prediction using machine learning. *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, 1–6 (2022).
 18. K. B. Sk, D. Roja, S. S. Priya, L. Dalavi, S. S. Vellela, V. Reddy, Coronary heart disease prediction and classification using hybrid machine learning algorithms. *Proc. Int. Conf. Innov. Data Commun. Technol. Appl. (ICIDCA)*, 1–7 (2023).
 19. D. Hassan, H. I. Hussein, M. M. Hassan, Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis. *Biomed. Signal Process. Control* 79, 104019 (2023).
 20. K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, T. Gonzales-Yanac, Multiple disease prediction using machine learning algorithms. *Mater. Today Proc.* 80, 3682–3685 (2023).
 21. C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, S. S. Ullah, Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors* 22, 7227 (2022).
 22. A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, N. Ullah, A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Inf. Syst.* 2022, 1410169 (2022).
 23. M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Avelaira-Mata, J. A. Benítez-Andrades, Heart disease risk prediction using deep learning techniques with feature augmentation. *Multimed. Tools Appl.* 82, 31759–31773 (2023).
 24. A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, S. N. Qasem, Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics* 14, 144 (2024).