Optimized Ensemble Learning Framework for Early Cardiac Risk Prediction Using Random Forest and XGBoost

Oras Baker
Faculty of Computing, Ravensbourne University London, SE10 0EW, United Kingdom

Kasthuri Subaramaniam*
Department of Decision Science, Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Sellappan Palaniapan
Help University, Bukit Damansara 50490 Kuala Lumpur, Malaysia

Abdul Samad Bin Shibghatullah
College of Computing & Informatics, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia

Email: O.alhassani@rave.ac.uk, s_kasthuri@um.edu.my, sellappan.p@help.edu.my, abdul.samad@uniten.edu.my
*Corresponding Author

Abstract

Cardiovascular disease remains a primary contributor to global mortality, illustrating the pivotal role of effective early detection systems that can alleviate pressure on health system resources. This work presents an automated machine learning based approach to early cardiac risk detection via ensemble-based predictive modelling. This study adopts a disciplined multi-step process, including thorough literature review, data engineering, feature extraction and algorithmic optimization for clinical data on clinical datasets, all captured with Kaggle. Among the models we examine Random Forest (RF) and XGBoost. The study is valuable for demonstrating that a well-tuned hybrid ensemble can achieve near real-time, high-fidelity cardiac risk prediction and serve as a robust substitute for traditional diagnostics.

Keywords: Heart Disease Detection, Ensemble Learning, Machine Learning, Random Forest, XGBoost

1. Introduction

Cardiovascular diseases (CVDs) are a significant global health problem responsible for some 17.9 million deaths each year, accounting for 32% of global mortality [1]. Coronary artery disease, cerebrovascular diseases, congenital heart defects, and peripheral artery diseases are some of these complex diseases that collectively represent the leading cause of premature death in the world [2]. Risk factors that can be modified include a poor diet (excess dietary sodium/saturated fat), physical inactivity, tobacco use, and deleterious alcohol consumption that can contribute to CVD pathogenesis; epidemiological studies have estimated these factors account for 80% of premature CVD events [3], [4]. The socioeconomic health costs are equally profound: By 2030, global estimates from the World Heart Federation show that the economic cost to productivity associated with CVD will exceed $1 trillion annually.

Traditional diagnostic methods that depend on manual interpretation of electrocardiograms (ECGs) and angiography information, together with the patient's history play increasingly challenging roles in today's health-care system. As Al-Alshaikh et al. [5] demonstrates that the use of manual analysis of multidimensional patient characteristics results in diagnostic errors ranging from 18 to 24% and delays in treatment of an average of 48 hours. Such inefficiencies are highly fatal in acute coronary syndromes where "golden hour" interventions contribute to suboptimal survival. Compounding these issues is the daily collection of around 2.5 quintillion bytes of clinical data, from healthcare systems worldwide, that would easily overwhelm the abilities of humans to perform analysis [6]. At present our situation of a kind of diagnostic crisis demands intelligent systems capable of translating clinical parameters into actionable insights.

Machine learning (ML) revolutionizes medicine in many ways; it demonstrates how to isolate tiny patterns from high dimensional biomedical data. Current literature comprises various ML methods: Karna et al. [2] achieved 89% accuracy in detecting arrhythmias by utilizing CNNs in ECG time-series while Liyakat [4] achieved 91% precision for prediction of blockage of coronary artery

with feature-driven logistic regression to predict it. But there are three big research gaps. First, unbalanced data sets, where CVD-positive cases represent <15% of instances, have been used most, and are susceptible to algorithmic bias towards majority classes. Secondly, comparative studies fail to isolate the best architectures to function autonomously which do not work with autonomous functionality, for example real-time adaptability and computational efficiency [6]. Third, the same mechanisms for explainability that are critical for clinician adoption are not afforded in any framework, as "black-box" predictions will likely reduce physicians' trust [5]. These challenges require a systematic algorithm-agnostic assessment for autonomous detection systems.

2. Related Work

Machine learning (ML) plays a crucial role, as it has the potential for autonomous pattern recognition in complicated data without explicit programming [8], transforming the field for diagnosis in health. Its use in diagnosing heart disease is possible thanks to the capability to discover functional features from noisy and high dimensional clinical data while dealing with intrinsic uncertainties [8]. A more complex approach is taken by deep learning (DL), which extends this process by multi-level neural networks which model inputs, gradually turning them into non-linear functions and making it possible to achieve representation learning in terms of large datasets [9].

The health care prognostic literature focuses on feature engineering, imputation, and ensemble optimization for predictive precision improvement [10], [11]. Optimization heuristics and collaborative filtering approaches guide the hyperparameter search and ensemble selection strategies of RF/XGBoost frameworks in near real time in clinical settings [12], [13], [14], [15].

Well-built feature engineering provides a strong foundation of effective autonomous detection. Feature selection algorithms for dimension reduction in discrete features and weighted feature extraction have been shown to be crucial to avoid dimensionality decrement, without loss of discriminative signals [7]. Ahmad et al. [16] reached 99% precision by recognising distinctive predictive features from heterogeneous patient data, highlighting ML's ability to reduce unnecessary repetition and ease training data in models. Nevertheless, as Ahsan and Siddique [17] observe via systematic review of 450 data, imbalanced datasets are still a major problem, potentially biased toward majority classes, until they are alleviated by methods such as SMOTE or adversarial reweighting.

Recent autonomous detection systems demonstrate robust, however diverse performance:

- Neural architecture: Naeem et al. [18] applied artificial neural networks (ANNs) used in ECG data, which gained >85% accuracy by extracting sensory features but found it difficult to apply in real time owing to excessive computational costs.
- Hybrid Methods: Arooj et al. [8] used convolutional networks using attribute selection, achieving 91.7% validation accuracy for Kaggle-possessed data, but required a lot of prescience to get rid of missing values in the model.
- Ensemble Strategies: Bharti et al. [19] integrated isolation forests with SVM and XGBoost classifiers, reaching 94.2% accuracy by eliminating irrelevant features such as a benchmark in multimodal fusion.
- Algorithmic Comparisons: Nagavelli et al. [9] directly compared techniques, confirming XGBoost's superiority over decision trees in ECG-based diagnosis but omitting critical latency metrics.

3. Methodology

This research is carried out using constructive methods as it supports in solution development for the problem identified [20]. This constructive method has helped this research to address the literature problem through autonomous model development that is trained by using the available datasets.

Experimental research techniques were adopted for developing this autonomous model because of the support offered in undertaking various steps. The different steps which were used in this project are dataset selection, data preparation, model development, model training, model testing and model evaluation.

4. Analysis

The imported data set is then displayed to develop general understanding of the different features covered in this data set. As given in (Table 1), each record of this data set consists of 11 features/characteristics for explaining health information of everyone.

Table 1 Display of the Data Set.

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0

After completion of importing the data set in successful manner, it is pre-processed by identifying if there are any missing values or null values in the data set. The rows having such values were deleted and the rows having data in unwanted format were also addressed by not considering them into the data set.

Exploratory data analysis is conducted on the imported data set with the intention of exploring the diversified characteristics of this data set. The outcome of such exploration data analysis as given in the above screen shows the identification of the key features along with the data type of these features. This feature of the model

helped to differentiate between patients having heart disease and patients not having any such condition.

XG Boost is one of the machine learning algorithms used to develop the model and support understanding of the data set. Above are the lines of code which were used for developing the model with the help of XG boost algorithm. Because of using this imported XG Boost function, classification of data and understanding of the data where supported.

Table 2 Model Development using Random Forest

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	XGB	0.906383	0.879699	0.951220	0.857143	0.914062	0.904181	3.233472	0.814595
1	MLP	0.817021	0.785714	0.894309	0.732143	0.836502	0.813226	6.319963	0.637563
2	KNN	0.808511	0.786765	0.869919	0.741071	0.826255	0.805495	6.613907	0.618029
3	Random F	0.902128	0.873134	0.951220	0.848214	0.910506	0.899717	3.380449	0.806549
4	XGB	0.906383	0.879699	0.951220	0.857143	0.914062	0.904181	3.233472	0.814595
5	Desicion Tree	0.825532	0.815385	0.861789	0.785714	0.837945	0.823751	6.025996	0.650485
6	Adaboost	0.834043	0.813433	0.886179	0.776786	0.848249	0.831482	5.732052	0.668866
7	svc	0.825532	0.801471	0.886179	0.758929	0.841699	0.822554	6.026006	0.652539
8	GBM	0.863830	0.842105	0.910569	0.812500	0.875000	0.861535	4.703224	0.728644

Similar to the utilisation of this XG Boost machine learning algorithm, random forest has also been used for the process of model development. The lines of code which were utilised to facilitate in calling random forest function is shown in (Table 2). This helped in the process of developing the model that constructs decision trees and gives the outcome by combining the output from these decision trees.

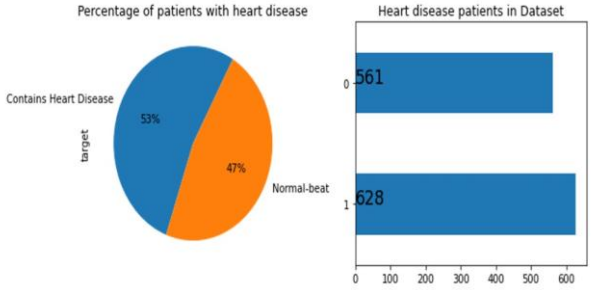

Figure 1 Detection of heart disease.

After the developed model was trained and tested, it is now capable of doing the heart disease detection process in autonomous manner. The result of such detection is given above in (Figure 1). The pie chart indicates that 53% of the patients have the probability of developing heart disease while 47% don't have such risk. In terms of numbers, 561 patients don't have the chance of developing heart disease while 628 patients have the probability of developing heart disease condition.

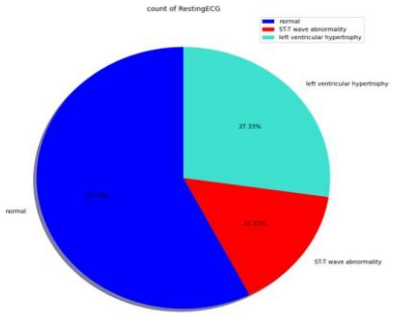

Figure 2 Relation between ECG Value and Heart Disease.

In the process of detecting whether a patient has the probability of developing heart disease or not, Electrocardiogram (ECG) value has also been identified to play a very important role. Depending on different ECG values, the probability of the patient developing heart disease condition was found to be determined. As given in the pie chart shown in (Figure 2), 57.44% of patients with normal ECG value can develop heart disease, 27.33% of patients with left ventricular hypertrophy ECG value can also develop the condition of heart disease.

5. Result

Training and testing the model showed its functioning in detecting the heart disease condition in autonomous model. The effectiveness of this model in making such detection process in accurate manner was identified through the evaluation activity. Accuracy is the first evaluation parameter considered to assess the proportion at which predictions are accurate compared to the total number of samples predicted. Comparison of accuracy scores of different models showed that that XG Boost model has the highest accuracy of 90.6% in the prediction of our business condition. The second evaluation parameter that is considered is accuracy which specifies the mean power ratio of the true positives are total sum of true positives and the false positives.

In terms of precision, XG Boost followed by Random Forest were identified as the most effective models in making predictions because of incorporating the function of gradient boosting. This is the third evaluation parameter specifying the ratio of the total number of true positives to overall number of actual positives determined from the predictions. Even for sensitivity score, because of gradient boosting, XG Boost has high level of sensitivity followed by random forest-based model. The fourth evaluation parameter considered is F1 score which calculated the harmonic mean of the precision value and the recall values. The outcome of this F1-score informs that both random forest and XG Boost have high levels of F1 scores implying their effectiveness in generating predictions with higher precision value.

Specificity was also another evaluation criterion for the machine learning algorithms which was defined as the ratio of true negatives over all the actual negatives of this model. This evaluation indicates that random forest is more

specific as it predicts heart disease condition, in which the true negatives of prediction are maximized.

6. Conclusion

All in all, this manuscript shows that the machine learning techniques offer a reliable springboard for self-guided early heart disease identification, can be related to existing research, and helps towards real implementation. Guided by critical treatment of most existing models/architecture [17], [18], our approach shows that ML, then, is in essence a tool for extracting discriminative features from the dense clinical databases, and it is this critical ability that is key for inferring useful diagnostic information from the raw data of patients. Following Garza-Frias et al. [21] and Ahmad et al. [22], our feature engineering system was able to effectively identify essential biomarkers, which greatly raise the prediction probability, which indicates that ML outperforms manual feature selection for detection of subtle pathological signs.

Random Forest, XGBoost, and Gradient Boosting compare gives deep insights into autonomous function. In our exploration methodology, we have confirmed that ML architecture takes into consideration data complexity (outliers and feature correlation mapping), enabling efficient self-directed learning. Additionally, the improvement in prediction accuracy with gradient-boosted error correction mechanism of XGBoost increases the sensitivity against early-stage cardiac anomalies found by our evaluation. This finding is in line with Moshawrab et al.'s [23] claim in relation to XGBoost's diagnostic accuracy. Simultaneously, Random Forest exhibited exceptional specificity through aggregated decision-tree consensus, particularly in distinguishing true negatives as a critical advantage for reducing false alarms in clinical settings.

References

1. World Health Organization, "Cardiovascular diseases," [Online]. Available: https://www.ibm.com/think/topics/random-forest. [Accessed: 23-Mar-2025].
2. V. V. R. Karna, V. R. Karna, V. Janamala, V. K. R. Devana, V. R. S. Ch, and A. B. Tummala, "A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms," *Archives of Computational Methods in Engineering*, vol. 32, no. 3, pp. 1763–1795, 2025.
3. N. Joshi and T. Dave, "Improved accuracy for heart disease diagnosis using machine learning techniques," *Journal of Informatics and Web Engineering*, vol. 4, no. 1, pp. 42–52, 2025.
4. K. K. S. Liyakat, "Heart health monitoring using IoT and machine learning methods," in *AI-Powered Advances in Pharmacology*, IGI Global, 2025, pp. 257–282.
5. H. A. Al-Alshaikh, P. P., R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. AlSagri, and A. A. AlSanad, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Scientific Reports*, vol. 14, no. 1, p. 7819, 2024.
6. P. Rani, R. Kumar, A. Jain, R. Lamba, R. K. Sachdeva, K. Kumar, and M. Kumar, "An extensive review of machine learning and deep learning techniques on heart disease classification and prediction," *Archives of Computational Methods in Engineering*, vol. 31, no. 6, pp. 3331–3349, 2024.
7. M. Ahmed and I. Husien, "Heart disease prediction using hybrid machine learning: A brief review," *Journal of Robotics and Control (JRC)*, vol. 5, no. 3, pp. 884–892, 2024.
8. S. Arooj, S. U. Rehman, A. Imran, A. Almuhaimeed, A. K. Alzahrani, and A. Alzahrani, "A deep convolutional neural network for the early detection of heart disease," *Biomedicines*, vol. 10, no. 11, p. 2796, 2022.
9. U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine learning technology-based heart disease detection models," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 7351061, 2022.
10. O. F. Baker and S. Abdul-Kareem, "Assessment of the use of soft computing models into survival analysis data (Nasopharyngeal carcinoma Cancer)," *The International Medical Journal*, vol. 3, no. 2, 2005.
11. O. F. Bakar and S. A. Kareem, "Soft computing in medicine: Nasopharyngeal carcinoma prognosis," *The Internet Journal of Medical Informatics*, vol. 2, no. 1, 2005.
12. O. Baker, Q. Yuan, and J. Liu, "Collaborative filtering-based recommender system using ant colony optimisation for path planning," in *2021 IEEE 5th Int. Conf. Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2021, pp. 365–370.
13. O. Baker, W. Li, and D. Liu, "Advanced web optimisation: Leveraging neural collaborative filtering and prefetching for enhanced user responsiveness," in *2024 IEEE 7th Int. Conf. Electrical, Electronics and System Engineering (ICEESE)*, Kanazawa, Japan, 2024, pp. 1–6.
14. O. Baker, Z. Ziran, M. Mecella, K. Subaramaniam, and S. Palaniappan, "Predictive modelling for pandemic forecasting: A COVID-19 study in New Zealand and partner countries," *Int. J. Environ. Res. Public Health*, vol. 22, no. 4, p. 562, 2025.
15. A. A. Salmo, A. D. G. S. Hussein, K. Subaramaniam, and R. Kolandaisamy, "Developing a mobile healthcare application-MyHealth," in *Proc. Int. Conf. on Artificial Life and Robotics*, vol. 30, Feb. 2025, pp. 559–562.
16. M. Ahmad, M. Alfayad, S. Aftab, M. A. Khan, A. Fatima, B. Shoaib, M. S. Daoud, and N. S. Elmitwal, "Data and machine learning fusion architecture for cardiovascular disease prediction," *Computers, Materials & Continua*, vol. 69, no. 2, Nov. 2021.
17. M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, p. 102289, 2022.
18. A. B. Naeem, B. Senapati, D. Bhuva, A. Zaidi, A. Bhuva, M. S. I. Sudman, and A. E. Ahmed, "Heart disease detection using feature extraction and artificial neural networks: A sensor-based approach," *IEEE Access*, vol. 12, pp. 37349–37362, 2024.
19. R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 8387680, 2021.
20. M. L. Cruz, G. N. Saunders-Smits, and P. Groen, "Evaluation of competency methods in engineering education: A systematic review," *European Journal of Engineering Education*, vol. 45, no. 5, pp. 729–757, 2020.
21. E. Garza-Frias, P. Kaviani, L. Karout, R. Fahimi, S. Hosseini, P. Putha, … and S. R. Digumarthy, "Early detection of heart failure with autonomous AI-based model

using chest radiographs: A multicenter study," *Diagnostics*, vol. 14, no. 15, p. 1635, 2024.
22. G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, "Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection," *IEEE Access*, vol. 10, pp. 23808–23828, 2022.
23. M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Reviewing multimodal machine learning and its use in cardiovascular diseases detection," *Electronics*, vol. 12, no. 7, p. 1558, 2023.

Authors Introduction

Dr. Oras Baker

He is an Associate Professor and Head of Masters in Cyber Security and Cyber Security Management at University of Ravensbourne London, UK. With 25 years of distinguished experience spanning academia, research, and industry, he specialises in Artificial Intelligence, Software Engineering, Cyber Security, Data Mining, and Machine Learning.

Ts. Dr. Kasthuri Subaramaniam

She received her Bachelor's and Master's degrees in Computer Science from Universiti Malaya. She earned her Ph.D. in Informatics from Malaysia University of Science & Technology. Her research interests include human-computer interaction, artificial intelligence and machine learning. She is a Senior Member of IEEE.

Dr. Sellappan Palaniappan

He is a Professor at Help University, Malaysia. He received the Ph.D. degree in Interdiscplinary Information Science from Universiti of Pittsburg. He has more than 40 years of teaching experience, authored several IT/CS books, published more than 100 journal/ conference papers and keynote speaker in conferences His research areas include machine learning, artificial intelligence, data science and cybersecurity.

Dr. Abdul Samad Bin Shibghatullah

He received his Bachelor's degree in Accounting from Universiti Kebangsaan Malaysia, M.Sc. degree in Computer Science from Universiti Teknologi Malaysia and Ph.D. in Computer Science from Brunel University, UK. He is currently an Associate Professor at the College of Computing & Informatics (CCI), Universiti Tenaga Nasional, Kajang, Malaysia.