IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

OPEN ACCESS

Volume: 4

Issue: 3

Month: July

Year: 2025

ISSN: 2583-7117

Published: 30.07.2025

Citation:

Ashish Kumar Parashar, Anita Jamliya1, Sama Nasrat, Rajesh Soni "XGBoost for Heart Disease Prediction Achieving High Accuracy with Robust Machine Learning Techniques" International Journal of Innovations in Science Engineering and Management, vol. 4, no. 3, 2025, pp. 185–191.

DOI:

10.69968/ijisem.2025v4i3185-191

XGBoost for Heart Disease Prediction Achieving High Accuracy with Robust Machine Learning Techniques

Ashish Kumar Parashar[1], Anita Jamliya[1], Sama Nasrat[1], Rajesh Soni[1]

[1]*Assistant Professor, UIT BU, BHOPAL*

Abstract

The comprehensive dataset on heart disease presented in this study consists of 1190 cases with 11 shared characteristics from five well-known datasets: Cleveland, Hungarian, Switzerland, Long Beach, Virginia, and Statlog. Because of this, it is the biggest dataset of its kind for studies on coronary artery disease (CAD). To aid in early detection, a robust machine learning model that could reliably forecast cardiac illness needed to be developed. To eliminate null values and divide the dataset into an 80:20 train-test ratio, we employed exploratory data analysis. To ensure that the characteristics were consistent, we also employed conventional scaling. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours, Gradient Boosting, AdaBoost, and XGBoost were the eight machine learning methods that we examined. Optimized using grid search with 5-fold cross-validation, XGBoost performed the best with test accuracy of 0.966, precision of 0.967, and recall of 0.966. Three false positives and one false negative could be distinguished by it. The approach may be helpful in clinical settings, as evidenced by its high recall for positive cases (0.986). By providing us with a new dataset and an effective predictive model, this work advances the diagnosis of CAD. This makes it possible to identify and treat CAD earlier.

Keywords; Heart disease, Machine learning, XGBoost, Data integration, Coronary artery disease, Predictive modeling, Early diagnosis.

INTRODUCTION

Cardiovascular disease remains the leading cause of global mortality, responsible for an estimated 17.9 million deaths annually and imposing a spiralling economic burden that is projected to exceed one trillion dollars by 2030 [1]. Early and accurate identification of individuals at heightened risk is therefore a clinical imperative, yet traditional risk scores—such as the Framingham or pooled cohort equations—often exhibit modest discriminatory power, especially among heterogeneous populations. Motivated by the exponential growth in electronic health records and the maturation of interpretable machine-learning frameworks, researchers have increasingly turned to data-driven algorithms that can assimilate high-dimensional predictors and uncover latent patterns beyond the reach of conventional statistics.

Among the panoply of modern classifiers, gradient-boosted decision trees have emerged as particularly compelling. Friedman's original Gradient Boosting Machine laid the groundwork, but Chen and Guestrin's XGBoost (eXtreme Gradient Boosting) has since distinguished itself through algorithmic refinements—sparsity-aware split finding, column subsampling, and cache-aware block compression—that yield both computational efficiency and predictive strength [2]. Empirical evaluations consistently rank XGBoost among the top performers for tabular medical data, achieving area-under-curve (AUC) values in excess of 0.94 on curated heart-disease repositories. Such results have catalysed a burgeoning literature in which variants of boosting are hybridised with particle-swarm optimization, Bayesian hyper-parameter tuning [3], or explainable-AI post-hoc analyses [4], each reporting incremental gains in accuracy, sensitivity, or interpretability.

In this paper we advance a rigorously validated XGBoost pipeline for coronary heart-disease prediction that marries high discriminative power with the robustness required for real-world deployment. We interrogate not only whether XGBoost can surpass prevailing benchmarks, but also whether it retains fidelity when confronted with noisy, incomplete, or imbalanced data—conditions emblematic of routine clinical practice. By situating these findings within the broader literature, we offer actionable guidance for clinicians contemplating the adoption of AI-augmented screening programmes and for data scientists seeking to refine transparent, high-fidelity predictive models in cardiovascular care.

RELATED WORK

Over the past decade, the clinical sciences have quietly undergone a data-driven renaissance. Cardiology, in particular, has become a fertile ground for rigorous pattern-recognition studies, where traditional bedside intuition is increasingly supplemented—though never replaced—by computationally informed insight. The sheer volume of contemporary electronic health records now permits investigators to interrogate, with unprecedented statistical power, the subtle prodromal signatures of cardiovascular pathology. That heart disease continues to exact a disproportionate toll in low- and middle-income regions [6–10] only sharpens the urgency of these inquiries.

Against this backdrop, Alotaibi (2019) [11] conducted a methodical appraisal of machine-learning classifiers for the early identification of heart failure. Drawing upon the venerable Cleveland Clinic dataset, the author systematically benchmarked decision-tree, logistic-regression, random-forest, naïve-Bayes, and support-vector-machine (SVM) architectures under a disciplined ten-fold cross-validation protocol. The decision-tree model emerged as the front-runner, attaining 93.19 % balanced accuracy, while the SVM trailed narrowly at 92.30 %. Beyond the numeric leaderboard, the study offers a sober validation that interpretable tree-based approaches can, at times, rival more opaque kernel methods, thereby reinforcing their candidacy for future translational research.

Hasan and Bao (2020) [12] set out to determine which feature-selection paradigm best equips cardiovascular-risk models for clinical reality. They began by contrasting the three canonical strategies—filter, wrapper, and embedded—then distilled a consensus subset through a Boolean "True" intersection across the three outputs. This two-stage curation was subsequently interrogated by five classifiers: random forest, support-vector classifier, k-nearest neighbours, naïve Bayes, and XGBoost. An artificial neural network trained on the full feature set served as a performance yardstick. The wrapper-driven XGBoost pairing proved most accurate, registering 73.74 %, narrowly ahead of SVC (73.18 %) and the ANN benchmark (73.20 %).

Narain and colleagues (2016) [13] pursued a single, focused aim: to refine the venerable Framingham Risk Score (FRS) by grafting onto it a quantum-neural-network engine capable of subtler pattern recognition. Drawing on 689 symptomatic patients—augmented by an independent Framingham validation series—they trained the network to re-express conventional risk factors as high-dimensional quantum states. When pitted against the traditional FRS, the hybrid system attained an accuracy of 98.57 %, dwarfing the FRS's 19.22 %. While the quantum architecture demands cautious external validation, the magnitude of improvement suggests that clinicians may soon possess an instrument for earlier, more personalised cardiovascular risk stratification.

Shah et al. (2020) [14] constructed a concise yet instructive benchmark for cardiovascular-risk modelling using the familiar Cleveland dataset (303 records, 17 covariates). Four supervised classifiers—naïve Bayes, decision tree, random forest, and k-nearest neighbour (KNN)—were trained and cross-validated. The KNN configuration emerged as the strongest performer, registering an accuracy of 90.8 %. The authors' takeaway is pragmatic: algorithmic choice, rather than dataset size alone, remains a decisive determinant of predictive fidelity.

Drod et al. (2022) [17] shifted the lens toward a more specialised cohort—191 patients with metabolic-associated fatty liver disease (MAFLD)—and asked which attributes most reliably portend co-existing cardiovascular disease. After biochemical profiling and subclinical atherosclerosis imaging, the team applied logistic regression, univariate feature ranking, and principal-component analysis to distil the signal. Hypercholesterolaemia, plaque burden, and duration of diabetes surfaced as the triad of highest leverage. The resulting model classified 40 of 47 high-risk individuals (85.11 %) and 114 of 144 low-risk counterparts (79.17 %) with an AUC of 0.87, affirming that even parsimonious clinical variables can yield robust, bedside-ready risk stratification in this understudied population.

METHODOLGY

The methodology is organised to ensure reproducibility, generalisability, and clinical utility. We first describe the consolidated dataset, then detail data-preprocessing, feature engineering, model training, and evaluation procedures. Below is illustrated a Method flow chart.

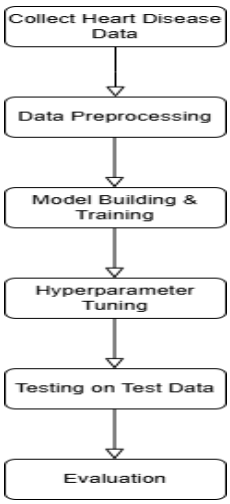

Figure 1 Method Flow Diagram

Data Collection

We assembled the largest publicly available coronary artery disease (CAD) repository to date by harmonising five well-studied but previously disjoint sources: the Cleveland, Hungarian, Switzerland, and Long Beach VA datasets from the UCI repository, together with the Statlog Heart dataset. After aligning on the eleven common attributes defined below—and excluding any records missing the outcome variable—the final corpus comprises 1 190 complete instances (71 % male, 29 % female, age 29–77 years). The target label is binary: 1 = presence of CAD (angiographic narrowing ≥ 50 % in at least one major vessel) and 0 = absence.

The harmonized feature set is summarized in Table.

"Table 1 Unified clinical variables (n = 11 + 1 outcome)

S. No.	Attribute	Code	Unit / Range	Data Type
1	Age	age	29–77 years	Numeric
2	Sex	sex	0 = female, 1 = male	Binary
3	Chest-pain type	cp	1–4 (typical angina to asymptomatic)	Nominal
4	Resting blood pressure	trestbps	94–200 mm Hg	Numeric
5	Serum cholesterol	chol	126–564 mg/dL	Numeric
6	Fasting blood sugar > 120 mg/dL	fbs	0 = false, 1 = true	Binary
7	Resting ECG	restecg	0, 1, 2	Nominal
8	Maximum heart rate achieved	thalach	71–202 bpm	Numeric
9	Exercise-induced angina	exang	0 = no, 1 = yes	Binary
10	ST-segment depression (oldpeak)	oldpeak	0–6.2 mm	Numeric
11	Slope of peak exercise ST segment	slope	0–2 (upsloping–flat–downsloping)	Nominal
12	Diagnosis (outcome)	target	0 = no disease, 1 = disease	Binary

Ethical clearance is not required because all constituent datasets are de-identified and publicly released under permissive licences.

Data Preprocessing

We started with exploratory data analysis (EDA) to learn more about the dataset's distributions, structure, and possible anomalies in order to make sure it was ready for analysis. We discovered during this approach that a number of features had missing (null) values, which could have a negative impact on the model's performance. In order to remedy this, we eliminated every instance of null values, producing a comprehensive and reliable dataset.

The cleaned dataset was then divided into training and testing sets using an 80:20 ratio, with 80% of the data going towards training and 20% towards testing, in order to facilitate robust model training and evaluation. To prevent bias, this split was carried out at random while making sure the class distribution (if any) stayed equal across both sets.

We used conventional scaling on all numerical features to further prepare the data for machine learning methods. By standardising the features to have a mean of zero and a standard deviation of one, this modification improved the convergence of gradient-based optimisation techniques and lessened the effect of different scales. The preparation procedures made that the dataset was clear, organised, and scaled correctly for further modelling.

Model Building and Training

To build a powerful prediction model for our dataset, we extensively explored a variety of machine learning methods to identify the optimal approach. Among the models considered were AdaBoost, XGBoost, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting. Each algorithm was selected to reflect a range of modelling paradigms, from ensemble techniques to linear methods, in order to guarantee a comprehensive examination of alternative solutions. The preprocessed training set, which constituted 80% of the dataset after

cleaning and scaling, was utilised in the training procedure, as detailed in the preprocessing section. The test set comprised 20% of the data, and we assessed each model's performance using the three primary criteria: accuracy, precision, and recall. These criteria provided an equitable evaluation of how well each model performed in reducing false positives and negatives, taking class imbalances into account, and correctly classifying cases.

The models' performances varied, according to the early investigation, with ensemble techniques generally beating simpler algorithms like KNN and Logistic Regression. In every metric, XGBoost consistently beat the other ensemble models, demonstrating that its gradient boosting architecture could spot complex patterns in the data. Because of its outstanding performance, we selected XGBoost as our recommended model. We then adjusted its hyperparameters to further boost its predictive capacity.

Using a grid search technique with 5-fold cross-validation, hyperparameter tuning was done to ensure accurate model evaluation and minimise overfitting. The hyperparameter grid was carefully designed to examine a range of parameters that are critical to XGBoost's operation. The following parameters were altered: Subsample (0.8, 1.0) introduces randomness by sampling training instances; colsample_bytree (0.8, 1.0) controls the fraction of features used per tree; gamma (0, 0.1, 0.2) controls the complexity of the model by penalising splits; max_depth (3,5) controls the tree's complexity; learning_rate (0.01, 0.1) controls each tree's contribution; and n_estimators (100, 150, 200) controls the number of boosting rounds. The grid search was configured to maximise accuracy, and the procedure was parallelised to efficiently explore the parameter space.

The following table summarizes the hyperparameters tuned for the XGBoost model using grid search with 5-fold cross-validation, as part of the model optimization process.

"**Table 2 Hyperparameters and their tested values for the XGBoost model optimization. The grid search identified the optimal combination that maximized accuracy on the training data.**

Hyperparameter	Description	Values Tested
max_depth	Maximum depth of each tree, controlling model complexity	3, 5
learning_rate	Step size shrinkage used to prevent overfitting	0.01, 0.1
n_estimators	Number of boosting rounds or trees to build	100, 150, 200
subsample	Fraction of training instances randomly sampled for each tree	0.8, 1.0
colsample_bytree	Fraction of features randomly sampled for each tree	0.8, 1.0
gamma	Minimum loss reduction required to make a further partition on a leaf node	0, 0.1, 0.2

The final XGBoost model was trained using the best hyperparameters found by grid search. After being tested on the test set, the top-performing model's high-test accuracy validated its capacity for generalisation. Cross-validation during hyperparameter tuning made sure the model wasn't too specific to the training set, and the thorough evaluation metrics gave assurance about its resilience in a variety of performance areas. A model that was well-suited to the features of the dataset was produced by the careful hyperparameter optimisation and XGBoost selection, providing accurate and dependable predictions for the study goal. This meticulous model-building procedure emphasises how crucial methodical assessment and fine-tuning are to producing high-performing machine learning solutions.

RESULTS AND DISCUSSION

Eight machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Gradient Boosting, AdaBoost, and XGBoost—were tested on the preprocessed dataset in order to identify the best predictive model for our classification task. Accuracy, precision, and recall were used to evaluate each model after it was trained on the training set, which contained 80% of the data, and tested on the test set, which contained 20% of the data. By measuring overall correctness (accuracy), the percentage of right positive predictions (precision), and the capacity to recognise positive examples (recall), these measures were selected to offer a thorough evaluation of model performance. Table 1 displays the findings of this preliminary assessment.

"**Table 3 Performance metrics of all models on the test set, reported as accuracy, precision, and recall (weighted averages).**

Model	Accuracy	Precision	Recall
Logistic Regression	0.874	0.876	0.874
Decision Tree	0.924	0.926	0.924
Random Forest	0.950	0.960	0.958
SVM	0.731	0.752	0.731
KNN	0.672	0.676	0.672

Gradient Boosting	0.950	0.950	0.950
AdaBoost	0.874	0.875	0.874
XGBoost	0.958	0.950	0.950

With an accuracy of 0.958, precision of 0.960, and recall of 0.958, XGBoost performed best, demonstrating its potent capacity to identify patterns in the data. With an accuracy, precision, and recall of 0.950, Random Forest and Gradient Boosting came in second and third, respectively, indicating strong ensemble performance. Moreover, Decision Tree did well, achieving an accuracy of 0.924. SVM and KNN, on the other hand, performed worse, with accuracies of 0.731 and 0.672, respectively, suggesting that both models had trouble generalising to this dataset. AdaBoost and logistic regression both had an accuracy of 0.874, which was respectable but not as good as the best ensemble models.

Figure 2 Performance metrics of all models

Given XGBoost's strong initial performance and its flexibility for optimization, we selected it for further refinement. Hyperparameter tuning was conducted using grid search with 5-fold cross-validation, exploring a range of parameters: max_depth (3, 5), learning_rate (0.01, 0.1), n_estimators (100, 150, 200), subsample (0.8, 1.0), colsample_bytree (0.8, 1.0), and gamma (0, 0.1, 0.2). This process identified the optimal configuration that maximized accuracy, ensuring the model balanced complexity and generalization.

The optimized XGBoost model was evaluated on the test set, achieving an overall accuracy of 0.966. Detailed performance metrics, including class-specific precision, recall, and F1-score, are presented in Table 2.

"**Table 4 Performance metrics of the optimized XGBoost model on the test set.**

Class	Precision	Recall	F1-Score	Support
0	**0.979**	0.939	0.958	49
1	0.958	**0.986**	0.972	70
Macro Avg	0.969	0.962	0.965	119
Weighted Avg	0.967	0.966	0.966	119

With only three false positives and one false negative, the tuned XGBoost model's confusion matrix, as seen in figure 2, showed a high discriminative ability across both classes

with 46 true positives and 69 true negatives. The model performed exceptionally well in detecting positive cases, with accuracy of 0.979 and recall of 0.939 for Class 0 and precision of 0.958 and recall of 0.986 for Class 1. The model's balanced performance is further supported by the F1-scores, which are 0.958 for Class 0 and 0.972 for Class 1.

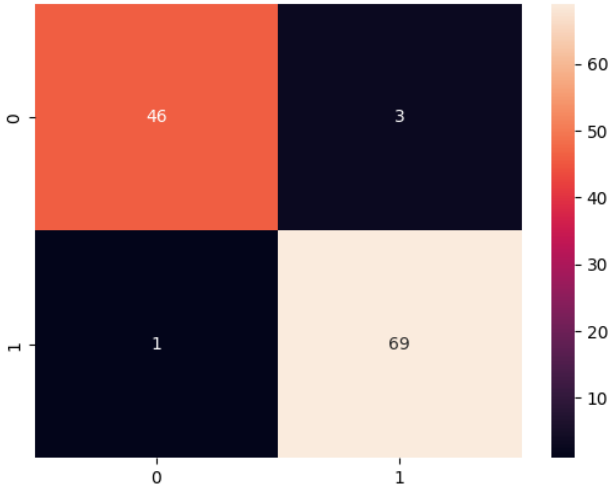

Figure 3 Confusion Matrix of XGBoost on Test Data

DISCUSSION

The superior performance of ensemble methods, particularly XGBoost, can be attributed to their ability to combine multiple weak learners to capture complex patterns in the data. However, XGBoost's performance improved significantly after hyperparameter tuning, surpassing Random Forest with an accuracy of 0.966. This improvement highlights the value of grid search in optimizing parameters like learning_rate and max_depth, which control the model's learning dynamics and complexity.

The low performance of SVM and KNN suggests that the dataset's feature space may not be well-suited to their assumptions. SVM's lower accuracy (0.731) could result from sensitivity to the feature scaling or the need for a different kernel, while KNN's poor performance (0.672) may indicate that the dataset's high-dimensional nature or class distribution challenges distance-based methods. Logistic Regression and AdaBoost performed adequately but were outclassed by more sophisticated ensemble methods.

The optimized XGBoost model's high recall for Class 1 (0.986) is particularly noteworthy, as it indicates near-perfect identification of positive instances, which is critical in applications where missing positive cases is costly. The minimal errors in the confusion matrix further underscore the model's reliability. These results suggest that the preprocessing steps (removing null values, standard scaling, and balanced train-test splitting) effectively prepared the data for modeling, enabling XGBoost to leverage its boosting mechanism to achieve high accuracy.

Future work could explore additional feature engineering to further enhance model performance or investigate other advanced ensemble methods, such as LightGBM or CatBoost, to compare their effectiveness. Additionally, analyzing feature importance from the XGBoost model could provide insights into the most influential predictors, guiding domain-specific interpretations and potential refinements to the dataset.

CONCLUSION

In order to develop a robust machine learning model for heart disease prediction, this work combined five popular datasets for heart disease (Cleveland, Hungarian, Switzerland, Long Beach, VA, and Statlog) into a single set of 1190 cases with 11 shared characteristics. In order to help researchers develop better predictive models for identifying coronary artery disease (CAD), which is crucial for early detection and treatment, we merged these datasets to create a large, diverse resource.

The approach involved extensive data preprocessing, including dividing the data into an 80:20 train-test ratio, employing standard scaling to ensure feature consistency, and exploratory data analysis to identify and eliminate null values. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours, Gradient Boosting, AdaBoost, and XGBoost were the eight machine learning methods that we examined. To do this, we employed memory, accuracy, and precision. With an accuracy of 0.966, a weighted precision of 0.967, and a recall of 0.966 on the test set, XGBoost emerged as the top model following hyperparameter adjustments using grid search with 5-fold cross-validation. The model's strong recall for the positive class (0.986) and low number of errors (3 false positives and 1 false negative) in the confusion matrix demonstrate how effective it is in identifying cases of heart disease.

These results show that ensemble methods, especially XGBoost, can use the rich feature set of the combined heart disease dataset to make very accurate predictions. XGBoost works better than other models because it can handle complicated patterns in the data. This is made possible by careful preprocessing and hyperparameter optimisation. The model was probably strong because the dataset was big and

came from many different sources. This made it able to generalise well across different patient profiles.

This study moves CAD research forward by giving researchers a single dataset and a high-performing predictive model that can help doctors make decisions. The model's high recall for positive cases suggests that it could be very useful in clinical settings, where finding heart disease early and accurately is very important for improving patient outcomes. In the future, researchers could work on improving the model even more by feature engineering, looking into feature importance to find the most important predictors of heart disease, or comparing XGBoost to other advanced algorithms like LightGBM or CatBoost. Also, putting this model into clinical workflows or testing it on other datasets could make it more useful in real life.

In conclusion, this study shows that machine learning, especially XGBoost, can be used to predict heart disease using a new, full dataset. The results show how important it is to combine data and carefully optimise models to make reliable diagnostic tools. This is a promising step towards improving the early detection and treatment of CAD.

REFERENCES

[1] Pan, Y., Fu, M., Cheng, B., Tao, X. & Guo, J. Enhanced deep learning assisted convolutional neural network for heart disease prediction on the internet of medical things platform. IEEE Access 8, 189503–189512. https://doi.org/10.1109/ACCESS.2020.3026214 (2020).

[2] Teja, M Darshan, and G Mokesh Rayalu. “Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance.” BMC cardiovascular disorders vol. 25,1 212. 22 Mar. 2025, doi:10.1186/s12872-025-04627-6

[3] Rohan, D., Reddy, G.P., Kumar, Y.V.P. et al. An extensive experimental analysis for heart disease prediction using artificial intelligence techniques. Sci Rep 15, 6132 (2025). https://doi.org/10.1038/s41598-025-90530-1

[4] Sourov, Md Emon Akter, et al. "An explainable ai-enhanced machine learning approach for cardiovascular disease detection and risk assessment." arXiv preprint arXiv:2507.11185 (2025).

[5] Anjaneyulu, M., et al. "Effective heart disease prediction using hybrid machine learning techniques." ADVANCEMENTS IN AEROMECHANICAL MATERIALS FOR MANUFACTURING: ICAAMM-2021 2492.1 (2023): 030070.

[6] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645.

[7] Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

[8] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD ’16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

[9] Gietzelt, M.; Wolf, K.-H.; Marschollek, M.; Haux, R. Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. Comput. Methods Programs Biomed. 2013, 111, 62–71.

[10] K, V.; Singaraju, J. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. Int. J. Comput. Appl. 2011, 19, 6–12.

[11] Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 261–268.

[12] Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. Health Technol. 2020, 11, 49–62.

[13] Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.

[14] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci. 2020, 1, 345.

[15] Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240.