Heart Disease Predictive Modeling with XGBoost and SMOTE-Driven Class Imbalance Mitigation

T. Roopa
School of Computing & Information Technology, REVA University, Bengaluru, Karnataka, India | SSIT, SSAHE University, Tumkur, India
troopa731@gmail.com (corresponding author)

Ganesh Dalappagari Ramanjinappa
School of Computing and Information Technology, REVA University, Bengaluru, Karnataka, India
ganesh.ramanjinappa@reva.edu.in

Received: 25 August 2025 | Revised: 14 September 2025, 16 September 2025, and 20 September 2025 | Accepted: 21 September 2025

 DOI: https://doi.org/10.48084/etasr.14301

ABSTRACT

Heart disease is one of the leading causes of death worldwide, highlighting the importance of early and precise diagnosis techniques. Clinical and demographic data can be analyzed using machine learning techniques to detect heart disease. This study combines XGBoost with the Synthetic Minority Oversampling Technique (SMOTE), introducing a strong prediction framework that handles class imbalance in the dataset. This study used 303 patient records with 13 clinical features from the UCI Heart Disease dataset, involving preprocessing, including addressing missing values, categorical variable one-hot encoding, and standardization of numeric features, followed by hyperparameter optimization using GridSearchCV. According to experimental findings, the model achieved an overall accuracy of 90%, with true positive and true negative counts of 96 and 66, respectively. The classification report shows good precision (0.87–0.93), recall (0.83–0.95), and F1-score (0.88–0.91), with low misclassification rates. Age, the type of chest pain, maximum heart rate, and cholesterol levels are highlighted as important predictors in the feature importance evaluation. According to the results, the XGBoost+SMOTE pipeline is very successful at classifying binary heart disease and can help with early therapeutic intervention techniques, which may lead to better patient outcomes.

Keywords-heart disease; machine learning; XGBoost; SMOTE; clinical diagnosis

I. INTRODUCTION

Nearly one-third of all fatalities globally are attributable to Cardiovascular Disease (CVD), which highlights the critical need for precise early diagnosis and prediction modeling. Accurate prediction of heart disease can prevent life threats, while incorrect prediction can be fatal [1]. In recent years, many studies have successfully used Machine Learning (ML) techniques to predict serious disease events using routine medical records [2]. ML is useful in diagnostic prediction and decision-making using health data [3]. The increasing size of medical datasets has made it a complicated task for practitioners to understand the complex feature relations and make disease predictions [4]. Among ML methods, XGBoost can be used to identify predictive factors because it can handle feature dependencies and mimic complex non-linear interactions [5]. Proper hyperparameter tuning is essential for any classifier's successful application [6]. In [4], various ML classification models were investigated using complete and reduced feature subsets [4].

Class imbalance is a common challenge in CVD datasets, since heart disease patients are underrepresented compared to healthy controls, potentially biasing predictions and reducing sensitivity toward minority classes. SMOTE-Edited Nearest Neighbor (SMOTE-ENN) is an oversampling strategy used to address this problem by processing unbalanced data to obtain roughly the same positive and negative classes [7, 8]. SMOTE improves the accuracy of model prediction, particularly for minority classes [9]. Web-based heart disease classification systems can help the public assess their risk of heart disease early, allowing them to take preventive action sooner [10]. In [11], the HDBN-XG algorithm assessed data quality, performed normalization using z-score, extracted features via the computational rough set method, and constructed feature subsets using the multi-objective artificial bee colony approach. The performance of heart disease prediction systems can be improved using algorithms such as KNN, ANN, RF, PCA, and GA [12]. In addition, the importance of each feature can be computed and visualized using Shapley Additive exPlanations (SHAP) [13].

This study used the UCI Heart Disease dataset, which includes 303 patient records with 13 attributes and a goal variable that indicates the presence or absence of heart disease, as the benchmark clinical dataset. It includes demographic information (age, sex), clinical information (serum cholesterol, resting blood pressure, maximum heart rate), and symptom indicators (type of chest pain, exercise-induced angina) [14].

In [15], GridSearchCV was used with five-fold cross-validation to optimize the hyperparameters of AdaBoost [15]. In [16], a hybrid sampling method combined SMOTE and Tomek links to mitigate the impact of unbalanced datasets on model performance. In [17], a novel SMOTE-based hybrid deep learning (SMOTE-HDL) network was proposed to predict the survival of patients with heart failure [17].

Traditional AI and ML models do not address data imbalance and lead to lower prediction accuracy. Compared to individual classifier models, ensemble methods perform better [18]. In [19], preprocessed data were fed into four popular classification algorithms, namely Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN). In [20], eXplainable AI (XAI) algorithms were investigated in predicting cardiovascular strokes.

According to the World Health Organization, heart disease causes about 30% of the total deaths and mainly occurs in individuals who are in their productive age [21]. The study in [22] emphasized the effectiveness of meticulously fine-tuning an XGBoost model for cardiovascular diseases. The integration of ensemble learning, imbalance mitigation, and XAI provides a promising framework for highly accurate, interpretable, and clinically actionable cardiovascular risk prediction systems. This study specifically focuses on developing, implementing, and evaluating an optimized SMOTE-XGBoost framework for early detection of heart disease, demonstrating the potential of combining classical ML techniques with modern AI interpretability methods.

II. PROPOSED SYSTEM

The UCI Heart Disease dataset served as the basis for this study. The dataset was preprocessed to address missing values, standardize features, and ensure data quality. Then, class imbalance was handled using SMOTE, and the processed dataset was used to train an XGBoost classifier to calculate the risk of heart disease. Figure 1 shows the overall flow of the proposed method.

A. Dataset Description

The UCI Heart Disease dataset is a benchmark in cardiovascular research and predictive modeling, including 13 clinical factors and a target variable that indicates the presence or absence of cardiac disease in each of the 303 patient records. These characteristics include demographics (age, sex), clinical data (maximum heart rate, resting blood pressure, serum cholesterol), as well as symptoms (exercise-induced angina, type of chest pain) [23]. Collectively, these features represent key cardiovascular risk factors and provide a solid basis for predictive modeling. Since the dataset is moderately imbalanced, with slightly more healthy cases, imbalance handling is essential for clinically reliable predictions.

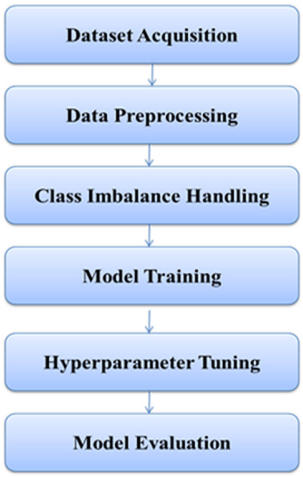

Fig. 1. The proposed method.

B. Data Preprocessing

Medical datasets often exhibit missing values, heterogeneous formats, and mixed data types, requiring preprocessing to ensure robust model performance. In this study, categorical attributes with missing values were imputed using the most frequent mode, while continuous features were imputed with the mean or median, depending on their distribution. Categorical variables (e.g., sex, chest pain type, exercise-induced angina) were transformed into q numeric format using one-hot encoding. Continuous attributes (e.g., age, cholesterol, resting blood pressure) were standardized using StandardScaler to achieve zero mean and unit variance, improving the training stability of XGBoost.

C. Addressing Class Imbalance

Class imbalance is a common problem in heart disease databases, where positive cases are underrepresented, reducing the sensitivity of disease identification. This study used SMOTE to address this issue. SMOTE enriches the representation while reducing overfitting by creating synthetic minority-class samples by interpolation between nearest neighbors, in contrast to straightforward duplication. By balancing the dataset, recall and F1-score improvements are achieved, which are critical to identifying high-risk patients.

D. Model Pipeline

The predictive workflow integrates preprocessing, oversampling, model training, and hyperparameter tuning. After imputation, encoding, and normalization,the dataset was balanced using SMOTE. Subsequently, an XGBoost classifier, which recognizes intricate and nonlinear relationships between cardiovascular risk factors, was trained using the processed data. Using GridSearchCV with 5-fold cross-validation, hyperparameter tuning was carried out by adjusting parameters such as gamma, subsample, colsample_bytree, learning_rate, max_depth, and n_estimators. This systematic optimization prevented underfitting or overfitting, thus improving the generalizability of the model to previously unseen patient data.

E. Model Evaluation

The model was assessed using classification metrics such as F1-score, recall, accuracy, and precision. Although accuracy provides an overall performance measure, the minority group's recall and F1-score were prioritized to maintain clinical credibility in the diagnosis of heart disease. Additionally, the discriminative ability of the model was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC), where values close to 1 indicate strong prediction performance. Finally, a feature importance analysis was conducted using the built-in XGBoost ranking, identifying important indicators that are in line with known cardiovascular risk factors, such as the type of chest discomfort, serum cholesterol, and resting blood pressure [11].

III. RESULTS AND DISCUSSION

A. Dataset Overview

The original UCI Heart Disease dataset contains 303 patient records with 13 features. After applying preprocessing steps such as one-hot encoding of categorical variables and SMOTE-based oversampling to balance class distribution, the dataset was expanded to 920 samples with 16 effective features. The most frequent value (mode) for each feature was imputed to fill in any missing values in the dataset. Although a feature warning regarding the downcasting of object-type arrays was observed throughout this procedure, the preprocessing procedures were successful in transforming the dataset into a format that was appropriate for training the model. An 80:20 split was used to divide the dataset into training and testing sets, yielding 736 training samples and 184 testing samples. This partitioning ensures sufficient data for effective model learning while preserving a representative distribution of classes in the test set.

B. Model Training and Hyperparameter Tuning

XGBoost is a gradient boosting decision tree method known for its excellent performance in classification challenges. GridSearchCV was applied to identify the optimal configuration, evaluating 729 different combinations of the following hyperparameters: *gamma*, *learning_rate*, *max_depth*, *n_estimators*, *subsample*, and *colsample_bytree*. During the tuning procedure, 5-fold cross-validation was used to eliminate the possibility of overfitting and ensure dependable model performance.

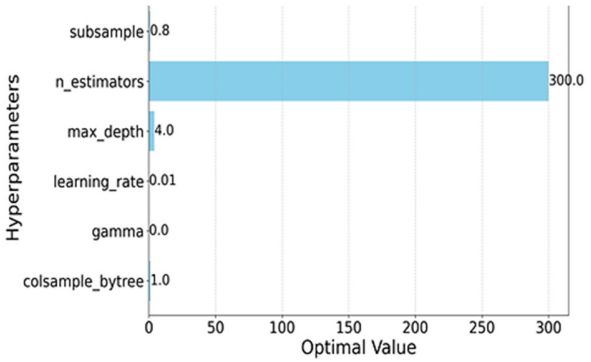

Fig. 2. The optimal hyperparameters for the XGBoost model.

Figure 2 shows the selected hyperparameters to balance accuracy and generalization. Using all features per tree (*colsample_bytree* = 1.0) and allowing unrestricted splits (*gamma* = 0) helps capture patterns effectively, while a small learning rate (0.01) with 300 estimators ensures gradual, stable learning. A *max_depth* of 4 prevents overly complex trees, and subsampling 80% of the data adds randomness, improving generalization and making the model reliable for disease prediction. These settings indicate that shallow trees with numerous boosting rounds and a small learning rate provided the best generalization for the dataset.

C. Performance Metrics

1) Accuracy and Classification Performance

With an overall accuracy of 90%, the XGBoost model showed excellent predictive performance on the testing dataset, demonstrating its ability to classify most cases correctly. Class-wise analysis further underscores the model's reliability in both categories. For the No Disease class (0), it achieved a precision of 0.93, a recall of 0.83, and an F1-score of 0.88, suggesting high precision in detecting healthy cases, though a few were incorrectly classified as diseased. With a precision of 0.87, a recall of 0.95, and an F1-score of 0.91 for the Disease class (1), the model demonstrated a great capacity to identify diseased instances with a low number of false negatives. The healthy class's support values, which show the actual sample counts per class, were 82, whereas the diseased class's were 102. Table I shows the classification results of the XGBoost model. Despite a small class imbalance, the model maintained balanced performance, achieving 90% accuracy with few false positives and negatives. Its reliability in distinguishing between healthy and diseased instances is demonstrated by strong class-level metrics, a weighted average of 0.90, and a macro-average F1-score of 0.89, suggesting its suitability for automated disease identification.

TABLE I. CLASSIFICATION PERFORMANCE OF THE XGBOOST MODEL

Class	Precision	Recall	F1-score	Support
No Disease(0)	0.93	0.83	0.88	82
Disease(1)	0.87	0.95	0.91	102
Macro Avg	0.90	0.89	0.89	184
Weighted Avg	0.90	0.90	0.90	184

2) Confusion Matrix Analysis

Figure 3 illustrates the classification performance of the XGBoost model based on the testing set, where 68 healthy cases were accurately classified as true negatives and 97 disease cases were appropriately identified as real positives. The algorithm misclassified fourteen healthy cases as diseased (false positives) and five disease cases as healthy (false negatives), which is a particularly important outcome—critical in healthcare settings as it reduces the likelihood of overlooking actual disease cases. Although a small number of healthy cases were misclassified as diseased, the model maintains a strong balance between sensitivity and specificity. This balance underscores its reliability for practical disease detection, ensuring accurate identification of true cases while minimizing unnecessary alerts.

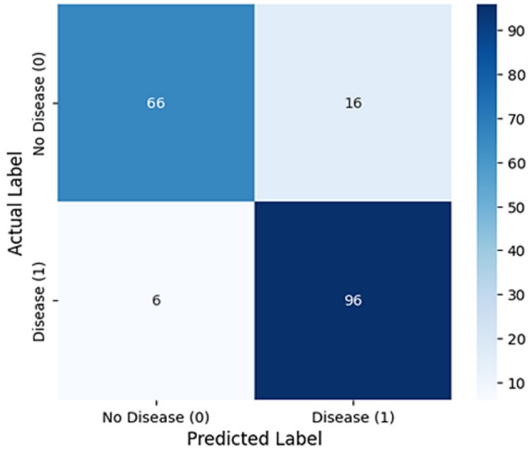

Fig. 3. Confusion matrix for the testing set.

3) *ROC Curve and AUC*

The ROC curve was used to evaluate the model's capacity to differentiate between healthy and disease instances. As shown in Figure 4, the model's AUC of 0.93 indicates that it performed significantly in discrimination. This indicates that the model consistently distinguishes positive from negative cases across varying classification thresholds, reinforcing its reliability for accurate disease detection.

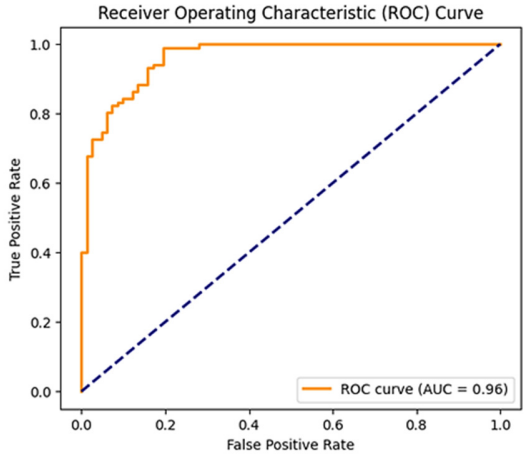

Fig. 4. The ROC curve of the XGBoost model.

4) *Feature Importance Analysis*

XGBoost inherently identifies the most important aspects in disease detection by ranking features according to their relevance to the prediction. Age, blood pressure, cholesterol, and other numerical characteristics had the greatest influence, while several categorical variables also had a significant impact following one-hot encoding. This analysis provides valuable, interpretable insights that can support clinical decision-making and help prioritize key patient indicators.

5) *Training vs. Testing Accuracy*

Figure 5 shows the impact of XGBoost's *n_estimators* option, which controls how many boosting rounds are used, to assess its effect on both training and testing accuracy. Adding more estimators did not result in significant overfitting because the model performed consistently throughout a range of boosting iterations. Furthermore, the model's ability to generalize to new data was demonstrated by the testing accuracy, which closely matched the training accuracy. This stability across cross-validation folds highlights the robustness of the model and confirms that the chosen boosting rounds provide a reliable balance between learning capacity and predictive generalization.

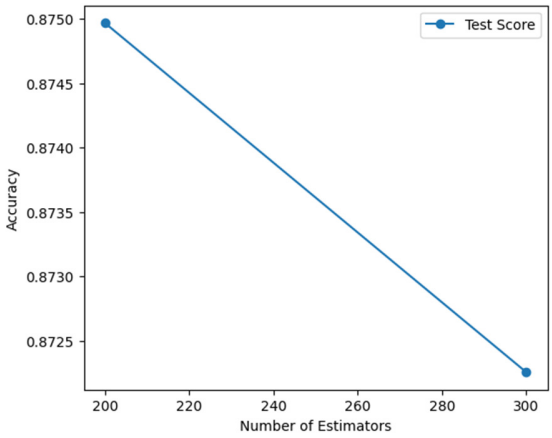

Fig. 5. Test accuracy with varying n_estimators.

D. *Comparison*

To evaluate the effectiveness of XGBoost, five ML algorithms were implemented with optimized parameters. XGBoost was configured with 100 estimators and a learning rate of 0.1, RF used 200 trees, SVM used the RBF kernel (C=1.0), Logistic Regression (LR) utilized L2 regularization, and KNN used k=5 based on validation results. All models underwent 5-fold cross-validation with an 80-20 train-test split. Table II compares the XGBoost model with RF, SVM, LR, and KNN. XGBoost outperformed RF (~87.5%), SVM (~85.3%), LR (~82.6%), and KNN (~80.4%), achieving the greatest accuracy of 90%. In accurately recognizing affirmative cases, it outperformed other models with a precision of 0.90 and a recall of 0.89. XGBoost showed balanced performance with the greatest F1-score of 0.90, avoiding false positives and false negatives while successfully controlling class imbalance.

TABLE II. COMPARISON OF VARIOUS MACHINE LEARNING ALGORITHMS' PERFORMANCE

Algorithm	Accuracy (%)	Precision (avg)	Recall (avg)	F1-score (avg)
XGBoost	90	0.90	0.89	0.90
RF	87.50	0.88	0.87	0.87
SVM (RBF kernel)	85.33	0.86	0.85	0.85
LR	82.60	0.84	0.82	0.83
KNN	80.43	0.81	0.80	0.80

From a performance interpretation perspective, RF demonstrated competitive results due to its ensemble nature but was slightly less optimized than XGBoost's gradient boosting approach. SVM with RBF kernel, while effective for complex boundaries, lagged in recall. LR and KNN with k=5 neighbors

showed comparatively lower accuracy, suggesting limited ability to capture complex feature interactions. Overall, these results clearly indicate that XGBoost provides the most reliable and robust performance for disease classification, as it achieved 90% accuracy with balanced performance across classes and minimal false negatives, reducing the risk of undetected disease cases. In addition, a ROC AUC of 0.93 confirmed its strong discriminative ability. The model has strong interpretability, identifying the most significant clinical characteristics, such as age, blood pressure, and cholesterol levels. Training and test performance remained stable across cross-validation and hyperparameter settings, demonstrating the model's reliability and robustness.

IV. CONCLUSION

This study demonstrates the effectiveness of an XGBoost-based predictive framework, augmented with SMOTE, for heart disease detection using a clinical dataset. The model achieved an accuracy of 90% and exhibited strong performance in identifying minority class cases, ensuring reliable detection of at-risk patients. Clinically significant variables, including age, the kind of chest discomfort, maximum heart rate, and cholesterol levels, were found to be important predictors, providing interpretable insights that can support early clinical decision-making and intervention. The robustness of the model, combined with its interpretability, suggests its practical applicability in supporting healthcare decision-making. Extending the model to predict multi-class heart disease severity and incorporating additional clinical parameters or lifestyle data could further improve its generalizability and predictive performance across diverse patient populations.

REFERENCES

[1] 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, Jan. 2021, Art. no. 8387680, https://doi.org/10.1155/2021/8387680.

[2] H. Khdair and N. M. Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021, https://doi.org/10.14569/IJACSA.2021.0120505.

[3] C. A. U. Hassan *et al.*, "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers," *Sensors*, vol. 22, no. 19, Sept. 2022, Art. no. 7227, https://doi.org/10.3390/s22197227.

[4] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, Nov. 2022, Art. no. 100060, https://doi.org/10.1016/j.health.2022.100060.

[5] Y. Xu *et al.*, "Predicting ICU Mortality in Rheumatic Heart Disease: Comparison of XGBoost and Logistic Regression," *Frontiers in Cardiovascular Medicine*, vol. 9, Feb. 2022, Art. no. 847206, https://doi.org/10.3389/fcvm.2022.847206.

[6] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, July 2022, https://doi.org/10.1016/j.jksuci.2020.10.013.

[7] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *School of Information, Shanxi University of Finance and Economics*, vol. 13, no. 10, Oct. 2022, Art. no. 475, https://doi.org/10.3390/info13100475.

[8] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Feb. 2022, https://doi.org/10.1155/2022/7351061.

[9] M. Aryuni, S. Adiarto, E. Miranda, E. D. Madyatmadja, V. D. S. Albert, and E. Sestomi, "Imbalanced Learning in Heart Disease Categorization: Improving Minority Class Prediction Accuracy Using the SMOTE Algorithm," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 23, no. 2, pp. 140–151, June 2023, https://doi.org/10.5391/IJFIS.2023.23.2.140.

[10] F. Novitasari, E. Haerani, A. Nazir, J. Jasril, and F. Insani, "Sistem Klasifikasi Penyakit Jantung Menggunakan Teknik Pendekatan SMOTE Pada Algoritma Modified K-Nearest Neighbor," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, June 2023, https://doi.org/10.47065/bits.v5i1.3610.

[11] K. Kalita, N. Ganesh, S. Jayalakshmi, J. S. Chohan, S. Mallik, and H. Qin, "Multi-Objective artificial bee colony optimized hybrid deep belief network and XGBoost algorithm for heart disease prediction," *Frontiers in Digital Health*, vol. 5, Nov. 2023, Art. no. 1279644, https://doi.org/10.3389/fdgth.2023.1279644.

[12] B. S. Peteti and D. Nandan, "Heart Disease Classification/Prediction: A Review," *Revue d'Intelligence Artificielle*, vol. 37, no. 2, pp. 347–377, Apr. 2023, https://doi.org/10.18280/ria.370213.

[13] G. Abdulsalam, S. Meshoul, and H. Shaiba, "Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 761–779, 2023, https://doi.org/10.32604/iasc.2023.032262.

[14] V. S. Devare, "Heart Disease Prediction Using Binary Classification," M.S. Thesis, California State University, USA, 2023.

[15] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *School of Electronics Engineering, VIT-AP University,* vol. 11, no. 4, Apr. 2023, Art. no. 1210, https://doi.org/10.3390/pr11041210.

[16] X. Liu, D. Li, and J. Zhao, "A Mortality Predicting Model for Heart Failure Patients Based on AdaBoost with Multi-kernel SVM," *Taiyuan University of Technology*, vol. 54, no. 5, 2023, https://doi.org/10.16355/j.tyut.1007-9432.2023.05.007.

[17] S. P. Barfungpa, L. Samantaray, and H. K. D. Sarma, "SMOTE-based adaptive coati kepler optimized hybrid deep network for predicting the survival of heart failure patients," *Multimedia Tools and Applications*, vol. 83, no. 24, pp. 65497–65524, Jan. 2024, https://doi.org/10.1007/s11042-023-18061-3.

[18] S. Naganjaneyulu, G. Akanksha, S. Shaheeda, and M. Sadhak, "HMLF_CDD_SSBM: A Hybrid Machine Learning Framework for Cardiovascular Disease Diagnosis Prediction Using the SMOTE Stacking Method," in *International Conference on Innovative Computing and Communications*, 2023, Delhi, India, pp. 571–585, https://doi.org/10.1007/978-981-99-3010-4_47.

[19] A. A. H. Alkurdi, "Enhancing Heart Disease Diagnosis Using Machine Learning Classifiers," *Fusion: Practice and Applications*, vol. 13, no. 1, pp. 08–18, 2023, https://doi.org/10.54216/FPA.130101.

[20] A. F. Tasnim *et al.*, "Explainable Machine Learning Algorithms to Predict Cardiovascular Strokes," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20131–20137, Feb. 2025, https://doi.org/10.48084/etasr.9152.

[21] A. N. Cahyani, J. Zeniarja, S. Winarno, R. T. E. Putri, and A. A. Maulani, "Heart Disease Classification Using Deep Neural Network with SMOTE Technique for Balancing Data," *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 1, Dec. 2023, Art. no. 0240108, https://doi.org/10.26877/asset.v6i1.17521.

[22] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2, Jan. 2024, Art. no. 144, https://doi.org/10.3390/diagnostics14020144.

[23] W. S. Andras Janosi, "Heart Disease." UCI Machine Learning Repository, 1989, https://doi.org/10.24432/C52P4X.