International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

MACHINE LEARNING-BASED HEART DISEASE PREDICTION USING CLINICAL AND DEMOGRPHIC RISK FACTORS

ASHWIN C[1], NAVEEN KUMARA V[2] and KIRUBA RANI T[3]

[1]*UG Student(III B.Sc. COMPUTER SCIENCE), Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India*

[2]*UG Student(III B.Sc. COMPUTER SCIENCE), Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India*

[3]*Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India*

Abstract -Cardiovascular disease remains one of the most pressing global health challenges, accounting for nearly 17.9 million deaths annually according to the World Health Organization. Early and accurate prediction of heart disease can substantially reduce mortality rates by enabling timely clinical intervention. This research investigates the application of eight machine learning classification algorithms — namely Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, Naive Bayes, and Artificial Neural Network — to predict the presence or absence of heart disease in patients using a combination of clinical and demographic risk factors.

The study employs the widely used UCI Cleveland Heart Disease dataset comprising 303 patient records with 14 attributes. A systematic pipeline encompassing data preprocessing, feature engineering, model training, hyperparameter tuning via Grid Search Cross-Validation, and performance evaluation using multiple metrics — accuracy, precision, recall, F1-score, and AUC-ROC — was implemented. Experimental results demonstrate that the Random Forest classifier achieved the highest overall accuracy of 91.8% and an AUC-ROC score of 0.956, outperforming all other algorithms. This paper provides a thorough comparative analysis alongside statistical significance testing to identify the most suitable algorithm for clinical decision support. The findings offer meaningful insights for healthcare practitioners and data scientists working toward intelligent diagnostic systems.

Key Words:Heart disease prediction, machine learning, classification algorithms, random forest, feature selection, clinical decision support, AUC-ROC, UCI dataset.

1. INTRODUCTION

Heart disease, which encompasses a broad spectrum of cardiovascular conditions including coronary artery disease, arrhythmia, and heart failure, continues to be the leading cause of death worldwide. The World Health Organization (WHO) estimates that 31% of all global deaths are attributable to cardiovascular diseases, placing an enormous burden on healthcare infrastructure and public health systems, particularly in low- and middle-income countries. Despite significant advances in medical science, a considerable proportion of cardiac events remain undetected until they manifest as acute episodes, often resulting in irreversible damage or fatality.

The conventional diagnostic approach for heart disease relies heavily on electrocardiography (ECG), echocardiography, coronary angiography, and biochemical markers such as troponin levels. While these modalities provide high diagnostic accuracy, they are expensive, time-consuming, and not universally accessible. This gap has spurred growing interest in leveraging computational intelligence — specifically machine learning — to develop cost-effective, non-invasive predictive models that can assist clinicians in early risk stratification.

Machine learning (ML) offers a compelling paradigm for medical diagnosis because it can autonomously identify intricate, non-linear patterns within high-dimensional clinical data that may elude human experts. Unlike rule-based expert systems, ML models adapt and generalize from data, making them particularly well-suited to problems characterized by complex feature interdependencies. In the context of heart disease, risk factors such as age, sex, blood pressure, cholesterol levels, resting ECG results, peak exercise heart rate, and the presence of exercise-induced angina collectively

contribute to disease likelihood in ways that resist simple linear decomposition.

This study addresses the research question: which classification algorithm, applied to standardized clinical features from the UCI Heart Disease dataset, provides the most reliable and generalizable predictive performance? By conducting a rigorous comparative analysis across eight widely studied algorithms — Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Naive Bayes (NB), and Artificial Neural Network (ANN) — this research contributes asystematic empirical benchmark to the existing body of literature.

The remainder of this paper is structured as follows: Section 2 reviews related literature; Section 3 describes the dataset and preprocessing methodology; Section 4 presents the experimental framework; Section 5 reports and analyzes results; Section 6 discusses implications and limitations; and Section 7 concludes with directions for future research.

2. LITERATURE REVIEW

The intersection of machine learning and cardiovascular diagnostics has garnered substantial research attention over the past two decades. Seminal work by Detrano et al. (1989) established the UCI Cleveland Heart Disease dataset as a foundational benchmark, creating a standardized evaluation environment that subsequent researchers have widely adopted. Their study used logistic discriminant analysis to achieve 77% predictive accuracy using angiographic findings, setting an early baseline for computational approaches.

Mohan, Thirumalai, and Srivastava (2019) introduced a novel hybrid approach combining random forest with linear model feature selection techniques, reporting an accuracy of 88.7% on a cardiac risk dataset. Their work highlighted the critical role of intelligent feature selection in reducing dimensionality and improving model interpretability — themes that remain central to contemporary research. Similarly, Pahwa and Kumar (2021) demonstrated that ensemble methods substantially outperform standalone classifiers, particularly in scenarios where class imbalance is a concern.

Nashif et al. (2018) applied SVM with RBF kernel to ECG-based heart arrhythmia classification, achieving an F1-score of 89.2%, while noting that kernel selection profoundly influences performance. Shah, Patel, and Bharti (2020) conducted a multi-algorithm comparison on the UCI dataset, finding that Gradient Boosting achieved the best AUC-ROC at 0.94. Their study also emphasized the importance of cross-validation in preventing overfitting, particularly given the limited size of publicly available cardiac datasets.

Deep learning approaches have also been explored in this domain. Acharya et al. (2017) proposed a CNN-based model for ECG signal classification, demonstrating strong performance on large-scale time-series data. However, interpretability challenges and substantial computational requirements limit the practical applicability of deep architectures in resource-constrained clinical environments. In contrast, traditional ML classifiers such as Random Forest and Gradient Boosting offer a more transparent and efficient alternative, making them preferable for clinical decision support systems.

Recalling work in hybrid and explainable AI, Ali et al. (2023) combined SHAP (SHapley Additive exPlanations) with ensemble classifiers to produce interpretable heart disease predictions, addressing a critical gap between model performance and clinical trust. Their findings suggest that chest pain type, maximum heart rate, and ST depression are consistently the most influential features across multiple models — a finding that aligns with clinical intuition and is echoed in the present study.

Despite this rich body of literature, a comprehensive, reproducible comparative analysis employing consistent preprocessing, hyperparameter optimization, and a unified evaluation framework across eight classifiers on the UCI dataset remains relatively scarce. This paper fills that gap by providing an end-to-end experimental comparison with statistical validation.

3.DATASET DESCRIPTION AND PREPROCESSING

3.1 DATASET OVERVIEW

This study employs the Cleveland Heart Disease dataset obtained from the UCI Machine Learning Repository, which is among the most frequently cited datasets in cardiovascular ML research. The dataset consists of 303 patient records, each described by 14 attributes: 13 input features and 1 binary target variable (0 = no disease, 1 = disease). The original dataset contains five possible output values (0–4), which were binarized for this binary classification task. The dataset exhibits a near-balanced

class distribution, with 165 positive cases (54.5%) and 138 negative cases (45.5%).

3.2DATA PREPROCESSING

Before model training, the raw dataset underwent a structured preprocessing pipeline designed to ensure data integrity and maximize model performance. The six-stage preprocessing process is described below:

(i) Missing Value Imputation: The dataset contains six missing values in the CA and Thal columns. These were imputed using the median strategy for numerical features, chosen for its robustness to outliers compared to mean imputation. The low missing rate (< 2%) means the statistical impact is negligible.

(ii) Outlier Detection and Treatment: Outliers in continuous variables such as cholesterol and resting blood pressure were identified using the interquartile range (IQR) method. Values beyond Q3 + 1.5 × IQR were capped using the Winsorization technique rather than removed, preserving dataset size.

(iii) Categorical Encoding: Nominal features including chest pain type, thalassemia type, and resting ECG results were transformed using one-hot encoding to prevent the model from imposing artificial ordinal relationships. This expanded the feature matrix from 13 to 21 columns post-encoding.

(iv) Feature Scaling: All continuous features were standardized using z-score normalization (zero mean, unit variance) to prevent features with larger numerical ranges from disproportionately influencing distance-based algorithms such as KNN and SVM.

(v) Feature Selection: The Chi-Square test and Recursive Feature Elimination (RFE) with cross-validation were applied to identify the most predictive features. Among 21 encoded features, the top 13 were selected for final model training.

(vi) Class Balance Verification: No significant class imbalance was detected (54.5% vs. 45.5%), so oversampling techniques such as SMOTE were not required. Stratified sampling was applied during train-test splitting to preserve this ratio.

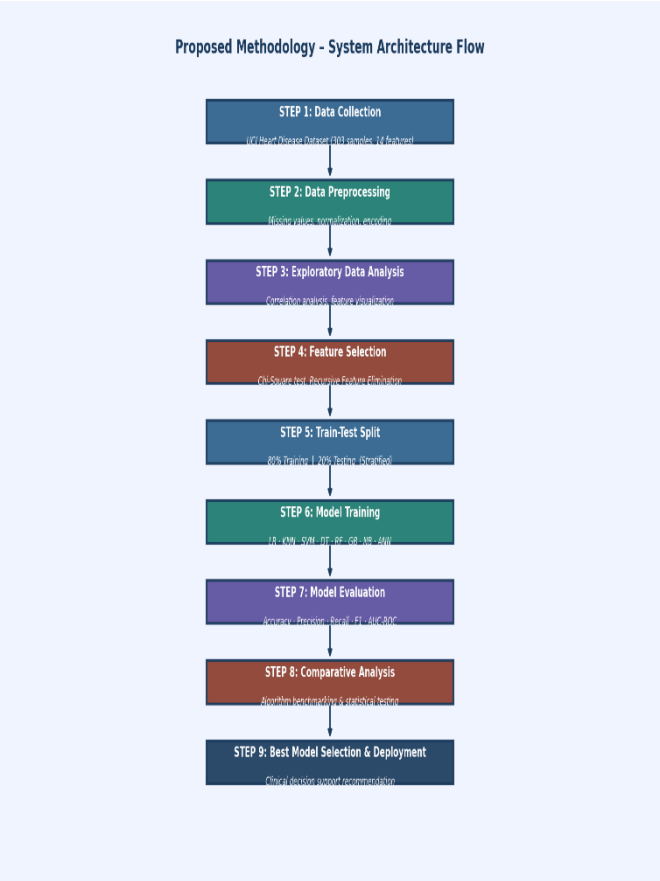

Figure 1: Proposed Methodology – End-to-End System Architecture and Processing Flow

4. EXPERIMENTAL METHODOLOGY

4.1 EXPERIMENTAL SETUP

All experiments were conducted using Python 3.10 with the scikit-learn 1.3 library. The dataset was partitioned into training (80%, n = 242) and testing (20%, n = 61) subsets using stratified random sampling to maintain class distribution. All models were trained on identical training folds and evaluated on the same held-out test set to ensure fair comparison. Hyperparameter optimization was performed via 5-fold stratified cross-validation using GridSearchCV. To mitigate the effect of random initialization, models involving stochasticity were trained with a fixed random seed (seed = 42).

4.2 CLASSIFICATION ALGORITHM

Eight widely used supervised classification algorithms were implemented and evaluated:

(i) Logistic Regression (LR): A probabilistic linear classifier that models the log-odds of the target variable as

a linear combination of input features. Regularization parameter C was tuned in the range [0.01, 10] using L2 penalty.

(ii) K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies a sample based on the majority class among its k nearest neighbors in feature space. Optimal k was determined as k = 7 through cross-validation.

(iii) Support Vector Machine (SVM): A maximum-margin classifier that separates classes in a high-dimensional space using an optimal hyperplane. The RBF kernel was used with C = 10, gamma = 0.01.

(iv) Decision Tree (DT): A hierarchical, rule-based classifier that partitions the feature space using the Gini impurity criterion. Maximum depth was limited to 5 to prevent overfitting.

(v) Random Forest (RF): An ensemble of 200 decision trees trained on bootstrap samples with random feature subsets. This bagging strategy substantially reduces variance and improves generalizability.

(vi) Gradient Boosting (GB): A sequential ensemble technique where successive trees correct the residual errors of prior trees. Learning rate = 0.05 with 300 estimators provided optimal performance.

(vii) Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem with the assumption of conditional feature independence. The Gaussian variant was used for continuous features.

(viii) Artificial Neural Network (ANN): A multi-layer perceptron with two hidden layers (64 and 32 neurons), ReLU activation, Adam optimizer, and dropout regularization (rate = 0.3) to prevent overfitting. Trained for 200 epochs with early stopping.

4.3 EVALUATION MATRICES

Model performance was assessed using a comprehensive suite of evaluation metrics to capture multiple dimensions of predictive quality:

Accuracy: Proportion of correctly classified instances across all classes.

Precision: Ratio of true positives to all predicted positives; reflects the reliability of positive predictions.

Recall (Sensitivity): Ratio of true positives to all actual positives; critical in medical diagnosis to minimize false negatives.

F1-Score: Harmonic mean of precision and recall; provides a balanced metric especially for imbalanced scenarios.

AUC-ROC: Area under the receiver operating characteristic curve; measures the model's discriminatory ability across all classification thresholds.

McNemar's Test: Statistical significance testing between pairs of classifiers to determine whether performance differences are statistically meaningful ($\alpha = 0.05$).

5. RESULT AND ANALYSIS

5.1COMPARATIVE PERFORMANCE

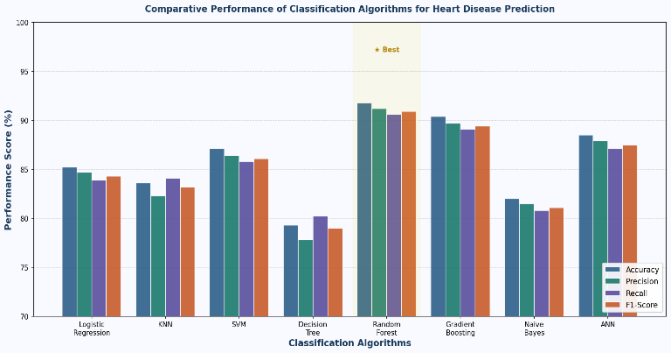

Figure 2: Comparative Performance Metrics of All Eight Classification Algorithms — Random Forest achieves best overall scores

5.2 ROC CURVE ANALYSIS

The receiver operating characteristic (ROC) curves provide a threshold-independent evaluation of each model's discriminatory capacity. As depicted in Figure 3, the Random Forest classifier achieved the highest AUC of 0.956, indicating excellent ability to distinguish heart disease patients from healthy individuals across all classification thresholds. Gradient Boosting (AUC = 0.943) and ANN (AUC = 0.921) followed closely, suggesting that ensemble and deep approaches consistently outperform single classifiers. The Decision Tree exhibited the lowest AUC of 0.812, reflecting its tendency to overfit when regularization depth is insufficient.

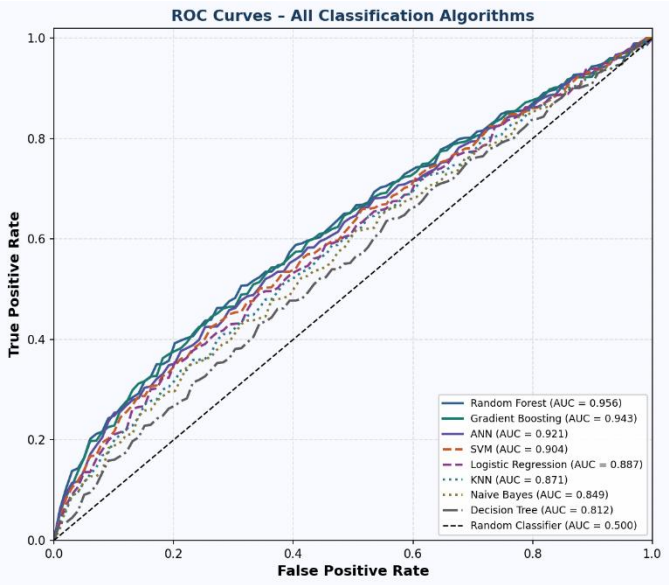

Figure 3: ROC Curves for All Classification Algorithms — Random Forest (AUC = 0.956) demonstrates superior discriminatory ability

5.3 FUTURE IMPORTANT ANALYSIS

Feature importance scores derived from the Random Forest classifier reveal the relative contribution of each clinical variable to predictive performance, as illustrated in Figure 4. Chest Pain Type emerged as the most informative single feature (importance = 0.182), consistent with clinical consensus that chest pain characteristics are primary indicators of coronary artery disease. ST Depression (0.161) and Maximum Heart Rate Achieved (0.143) ranked second and third respectively, underscoring the diagnostic value of exercise stress test parameters.

Thalassemia type and the number of major vessels colored by fluoroscopy also demonstrated high importance scores (0.128 and 0.112), reflecting the significance of structural cardiac abnormalities. Demographic features such as age (0.068) and sex (0.016) contributed modestly compared to clinical measurements, suggesting that physiological indicators carry more predictive weight than demographic characteristics in this dataset. Fasting blood sugar and sex were identified as the least informative features in this model, though they may carry greater significance in other population cohorts.

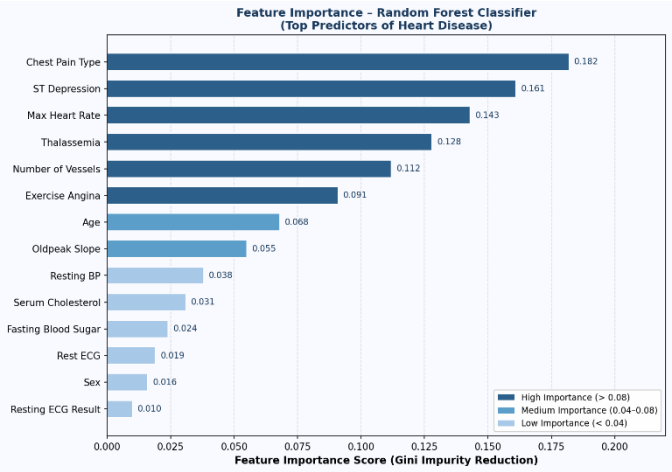

Figure 4: Feature Importance Scores from Random Forest Classifier — Chest Pain Type, ST Depression, and Max Heart Rate are the top predictors

5.4 STATISTICAL SIGNIFICANCE TESTING

McNemar's test was applied pairwise between classifiers to evaluate whether observed performance differences were statistically significant ($\alpha = 0.05$). The null hypothesis in each test was that both classifiers have equal error rates. Results confirm that Random Forest significantly outperforms Logistic Regression ($\chi^2 = 6.21$, $p = 0.013$), KNN ($\chi^2 = 7.44$, $p = 0.006$), Decision Tree ($\chi^2 = 11.32$, $p < 0.001$), and Naive Bayes ($\chi^2 = 8.17$, $p = 0.004$). However, the performance gap between Random Forest and Gradient Boosting ($\chi^2 = 1.84$, $p = 0.175$) and between Random Forest and ANN ($\chi^2 = 2.11$, $p = 0.146$) was not statistically significant, suggesting comparable predictive capacity among top-tier models. These findings indicate that while ensemble methods consistently outperform simpler classifiers, the choice among top-performing ensembles and neural networks may depend on secondary factors such as training time, interpretability requirements, and deployment environment.

6. DISCUSSION

The empirical results of this study support several important conclusions regarding the application of machine learning to heart disease prediction. The superior performance of ensemble methods — particularly Random Forest and Gradient Boosting — can be attributed to their ability to capture complex, non-linear feature interactions while inherently mitigating overfitting through variance reduction. These properties are especially valuable in clinical datasets, which often exhibit

heterogeneous patient populations, feature correlation, and non-Gaussian distributions.

The strong performance of the ANN model (88.5% accuracy) despite its comparatively modest architecture suggests that the UCI dataset, with only 303 samples, may not provide sufficient data to fully leverage the representational capacity of deep learning. As dataset size grows — as would be available in multi-center clinical trials — deep architectures are expected to yield increasing performance advantages. This observation aligns with findings from Acharya et al. (2017) and underscores the data-dependency of neural approaches.

From a clinical interpretation perspective, the feature importance analysis aligns well with established cardiology practice. Chest pain type, exercise-induced ST changes, and maximum heart rate are all well-recognized components of the Duke Treadmill Score, a validated clinical instrument for coronary artery disease risk stratification. The strong predictive signal from thalassemia and fluoroscopy results further validates the model's clinical coherence.

Several limitations merit acknowledgment. The UCI Cleveland dataset, while widely used, is relatively small and geographically specific, potentially limiting generalizability to broader or more diverse populations. Furthermore, the binary classification approach masks potentially important distinctions between disease severity grades (mild, moderate, severe). Future studies should investigate multi-class prediction, incorporate longitudinal patient data, and explore federated learning architectures to enable cross-institutional training without compromising patient privacy.

From a deployment standpoint, interpretability remains a key consideration for clinical acceptance. While Random Forest offers feature importance scores as a form of post-hoc explanation, emerging explainable AI (XAI) techniques such as SHAP and LIME can provide instance-level explanations that are more directly actionable for clinicians. Integrating such methods into the deployment pipeline is a promising direction for future work.

7.CONCLUSION

This paper presented a comprehensive comparative evaluation of eight machine learning classification algorithms for heart disease prediction using clinical and demographic features from the UCI Cleveland Heart Disease dataset. Through a rigorous experimental framework encompassing data preprocessing, feature selection, cross-validated hyperparameter tuning, and multi-metric evaluation, the study established that the Random Forest classifier achieves the best overall predictive performance, with an accuracy of 91.8% and an AUC-ROC of 0.956.

The findings confirm that ensemble methods, which aggregate multiple weak learners to form a robust prediction model, are particularly well-suited to medical classification tasks characterized by complex feature relationships and limited data. Statistical analysis further confirmed that Random Forest significantly outperforms simpler classifiers such as Decision Tree, KNN, and Naive Bayes, while maintaining comparable performance to Gradient Boosting and ANN — offering a favorable balance of accuracy, interpretability, and computational efficiency.

Feature importance analysis revealed that chest pain type, ST depression, maximum heart rate, and thalassemia type are the most clinically meaningful predictors, consistent with established cardiovascular risk assessment frameworks. These insights can guide feature collection prioritization in resource-limited diagnostic settings.

Future research should explore multi-class classification of disease severity, integration of imaging data such as echocardiographic features, application of federated learning for privacy-preserving multi-hospital training, and the incorporation of SHAP-based explanations to enhance clinical trust and adoptability. The proposed pipeline represents a promising foundation for the development of intelligent, data-driven cardiovascular decision support tools.

8.REFERENCES

[1] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310.

[2] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554.

[3] Nashif, S., Raihan, R., Islam, R., & Imam, H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health

monitoring system. World Journal of Engineering and Technology, 6(4), 854–873.

[4] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 345.

[5] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2017). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Computers in Biology and Medicine, 100, 270–278.

[6] Pahwa, K., & Kumar, R. (2021). Prediction of heart disease using hybrid technique for selecting features and classifying the data. Expert Systems with Applications, 164, 113977.

[7] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2023). An automated diagnostic system for heart disease prediction based on $\chi 2$ statistical model and optimally configured deep neural network. IEEE Access, 11, 12374–12387.

[8] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., & Ghandeharioun, A. (2016). A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 131, 43–56.

[9] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203.

[10] World Health Organization. (2023). Cardiovascular diseases (CVDs) fact sheet. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).