

An Optimal Machine Learning Framework for Early Detection of Heart Disease

Ibeh, C.H.¹, Chikwendu, C.R.^{2*}, Unaegbu, E.N.³

Department of Mathematics, Nnamdi Azikiwe University, Awka, Nigeria

*Corresponding Author

DOI: <https://doi.org/10.51584/IJRIAS.2026.11080198>

Received: 11 September 2026; Accepted: 16 September 2026; Published: 27 September 2026

ABSTRACT

Heart disease remains a major cause of mortality worldwide, creating a need for reliable early-risk identification. This study developed a cost-sensitive machine learning decision framework using 319,795 observations from the 2020 Behavioral Risk Factor Surveillance System (BRFSS) dataset. The cleaned dataset contained 18 predictor variables and a binary heart-disease outcome, with heart disease reported by 8.6% of respondents. Logistic Regression, Decision Tree, and Random Forest models were trained after Synthetic Minority Over sampling Technique (SMOTE) was applied to the training set. Performance was evaluated using accuracy, recall, precision, F1-score, and Matthews Correlation Coefficient (MCC). Random Forest achieved the highest accuracy (85.19%) but had low recall (30.56%). Logistic Regression achieved 69.78% recall, whereas the Decision Tree achieved 72.70% accuracy, 69.06% recall, 30.67% F1score, and the highest MCC (0.2581). Confusion-matrix analysis showed that the Decision Tree detected 3,862 heart-disease cases with 15,729 false positives, while Logistic Regression detected 3,902 cases but produced 17,430 false positives. A Bayesian cost-sensitive analysis assigned a false-negative cost ten times the false-positive cost and identified an optimal Decision Tree threshold of 0.36. The findings indicate that predictive accuracy alone is insufficient for imbalanced clinical screening and that an interpretable model combined with cost-sensitive threshold selection can provide a more balanced decision-support framework.

Keywords: heart disease; machine learning; decision tree; random forest; logistic regression; SMOTE; Bayesian decision theory; BRFSS; heart disease prediction; cost-minimization curve; Recall; MCC.

INTRODUCTION

Cardiovascular disease remains a major global health burden. The World Health Organization (WHO) reports that cardiovascular diseases account for approximately 17.9 million deaths annually, underscoring the importance of prevention and timely detection (World Health Organization, 2017). Early identification is important because risk information can support further clinical assessment and preventive intervention. However, cardiovascular risk assessment may involve multiple heterogeneous clinical and behavioural variables, making consistent evaluation difficult at population scale. Machine learning (ML) provides methods for learning predictive relationships from such multidimensional data and has increasingly been applied to cardiovascular medicine (Krittanawong et al., 2019; Rajkomar et al., 2019).

Different ML algorithms provide different balances of predictive performance and interpretability. Logistic Regression offers a transparent probabilistic baseline, Decision Trees provide rule-based partitioning, and Random Forests combine multiple trees to reduce variance (Breiman, 2001; Rudin, 2019). In health applications, however, a model that maximizes overall accuracy may still perform poorly for the minority disease class. This is particularly important when the cost of a missed diagnosis is substantially greater than the cost of an additional clinical assessment. Cost-sensitive learning and decision analysis therefore provide a principled way of incorporating unequal error consequences into classification decisions (Elkan, 2001; Vickers and Elkin, 2006).

Class imbalance is another central challenge in this study. When positive cases are much less frequent than negative cases, a classifier may achieve apparently high accuracy while failing to identify many patients with disease. The Synthetic Minority Over-sampling Technique (SMOTE) addresses this problem by generating synthetic minority observations between neighbouring minority-class examples (Chawla et al., 2002). For imbalanced binary classification, MCC is also informative because it incorporates all four confusion-matrix components and is less susceptible to misleading accuracy values (Chicco and Jurman, 2020).

The present study therefore compares Logistic Regression, Decision Tree, and Random Forest models using the 2020 CDC BRFSS heart-disease dataset and integrates predictive modelling with Bayesian cost-sensitive threshold optimization. The objective is to identify a model that provides an appropriate balance among sensitivity, overall performance, interpretability, and the consequences of diagnostic errors.

METHODS

Data Source and Preprocessing

The study used the cleaned 2020 Behavioral Risk Factor Surveillance System dataset released by the Centers for Disease Control and Prevention. The dataset contains 319,795 observations and 18 variables describing demographic, behavioural, and health characteristics. The response variable, HeartDisease, is binary, with 0 representing no heart disease and 1 representing heart disease. Predictor variables include BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, and SkinCancer.

Categorical variables were transformed into numerical representations suitable for machine learning. Binary variables were label encoded, while multi-category variables such as Age Category and Race were one-hot encoded. Continuous variables were normalized to a common scale using Min-Max transformation:

$$x_{scaled} = \frac{x - \min(X)}{\max(X) - \min(x)} \quad (2.1)$$

The target distribution was substantially imbalanced, with heart disease accounting for approximately 8.6% of the observations. SMOTE was therefore applied to the training data to generate synthetic minority-class observations. For a minority observation x_i and one of its selected nearest neighbours x_{zi} , a synthetic observation was generated as:

$$x_{syn} = x_i + \delta(x_{zi} - x_i) \quad 0 \leq \delta \leq 1 \quad (2.2)$$

The complete dataset was divided into training and testing subsets using an 80/20 split. The training set contained 255,836 observations and the unseen test set contained 63,959 observations. SMOTE was applied within the training process so that evaluation remained based on unseen test observations. The probability calibration techniques (such as Platt Scaling or Isotonic Regression) was not used to correct any probability distortions introduced by SMOTE in this study, instead we explicitly optimized the decision threshold using Bayesian Cost-Sensitive analysis to align predicted probabilities with clinical cost priorities.

Model Selection and Mathematical Formulation

Logistic Regression

Logistic Regression estimates the probability of heart disease using the logistic function:

$$P_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)]} \quad (2.3)$$

It provides a probabilistic and relatively interpretable baseline for binary classification.

Decision Tree

The Decision Tree partitions the feature space into increasingly homogeneous regions. The thesis formulation uses Gini impurity:

$$G(s) = 1 - \sum_{i=1}^k P_i^2 \quad (2.4)$$

where p_i is the proportion of observations belonging to class k in node S .

Random Forest

Random Forest combines predictions from multiple bootstrap-trained decision trees. The ensemble reduces variance through aggregation and random feature selection. Its greater complexity can improve predictive accuracy but reduces direct interpretability compared with a single tree.

$$f(x) = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (2.5)$$

Bayesian Decision Analysis

The study treats a false negative as more costly than a false positive. The implementation assigns a false-negative cost of 10 and a false-positive cost of 1. For a threshold τ , the total expected cost is represented by:

$$TER(\tau) = P(Y = 1)(1 - TPR(\tau))c_{01} + P(Y = 0)FPR(\tau)c_{10} \quad (2.6)$$

The computational analysis identifies $\tau = 0.36$ as the threshold producing the minimum total weighted clinical cost. This shifts the decision boundary below the conventional 0.50 threshold in order to favour earlier identification of potential heart disease cases.

Performance Evaluation

The models were evaluated on the unseen test set using accuracy, recall(sensitivity), precision, F1-score and MCC. Recall is especially important because it measures the proportion of actual heart disease cases that are correctly detected. MCC was also retained because it is informative under substantial class imbalance.

$$sensitivity = \frac{TP}{TP + FN} \quad (2.7)$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (2.8)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2.9)$$

RESULTS AND DISCUSSION

Exploratory Analysis

Exploratory analysis showed that the dataset contained a pronounced imbalance between respondents with and without heart disease. The analysis also indicated that cardiac risk was associated with multiple demographic, behavioural, and health-related factors rather than a single predictor. Age showed a particularly strong relationship with heart disease prevalence, with the prevalence increasing markedly across older age categories. Stroke, diabetes, and difficulty walking were also associated with disproportionately higher levels of heart disease within their respective groups.

Table 1: Exploratory Summary Statistics for Continuous Health Metrics

Feature	Mean	Std. Dev.	Minimum	Maximum
BMI	28.32	6.35	12.02	94.85
Physical Health	3.37	7.91	0.00	30.00
Mental Health	3.89	7.95	0.00	30.00
Sleep Time	7.09	1.43	1.00	24.00

The dataset contained 319,795 observations, of which 8.6% represented heart disease and 91.4% represented no heart disease. Continuous-variable summaries were: BMI, mean 28.32 (SD 6.35); Physical Health, mean 3.37 (SD 7.91); Mental Health, mean 3.89 (SD 7.95); and Sleep Time, mean 7.09 (SD 1.43).

Comparative Model Performance

The three models were evaluated on the unseen test set of 63,959 observations. Table 2 summarizes the resulting performance measures.

Table 2: Consolidated performance metrics for heart disease prediction models.

Model	Accuracy	Recall	Precision	F1-Score	MCC
Logistic Regression	0.7011	0.6978	0.1829	0.2899	0.2391
Decision Tree	0.7270	0.6906	0.1971	0.3067	0.2581
Random Forest	0.8519	0.3056	0.2341	0.2651	0.1864

Random Forest achieved the highest overall accuracy at 85.19%. However, its recall was only 30.56%, indicating that a large proportion of actual heart disease cases were not detected at the evaluated operating point. In contrast, Logistic Regression achieved a recall of 69.78%, demonstrating stronger sensitivity but lower overall accuracy and precision. The Decision Tree achieved 72.70% accuracy and 69.06% recall, while also recording the highest F1-score (0.3067) and MCC (0.2581).

The results demonstrate that the most accurate model is not automatically the most appropriate model for a screening-oriented task. Random Forest performed best under the global accuracy criterion but poorly under the sensitivity criterion. The Decision Tree offered the strongest combined performance across the imbalance-aware metrics used in this study. Its interpretable structure also provides an advantage when model outputs need to be examined alongside clinical judgement.

Confusion-Matrix Analysis

The following confusion matrices present the true negative(TN), false negative(FN), true positive(TP), false positive(FP) outcome for the three models on the unseen test set. They make the error patterns underlying the

consolidated performance metrics explicit. False negatives represent missed heart disease cases, while False positives represent healthy respondents incorrectly flagged as positive.

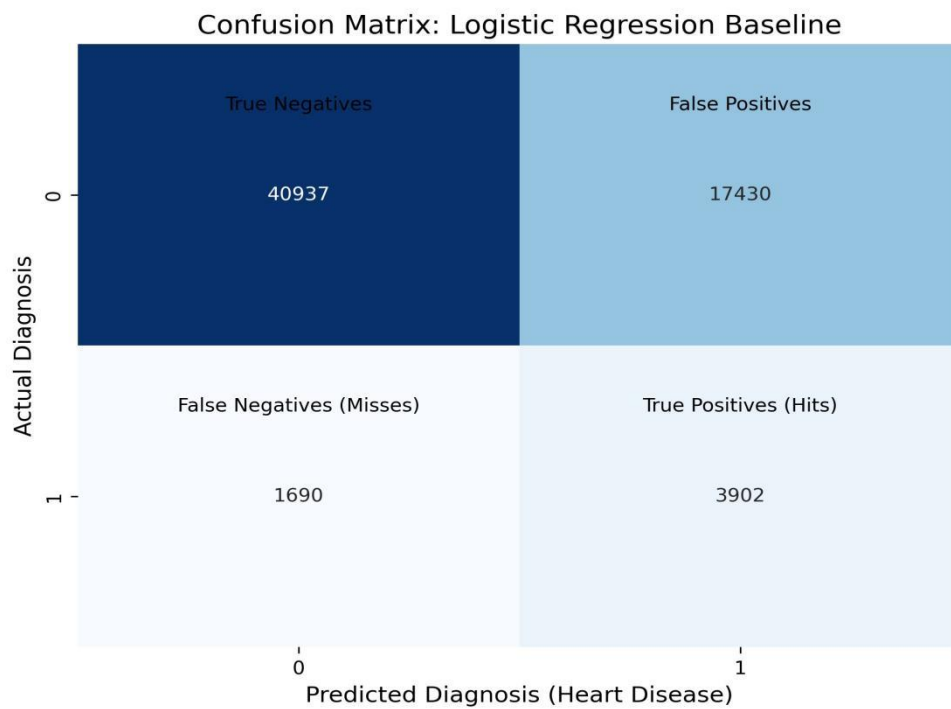


Figure 1: Confusion Matrix for the Logistic Regression Baseline

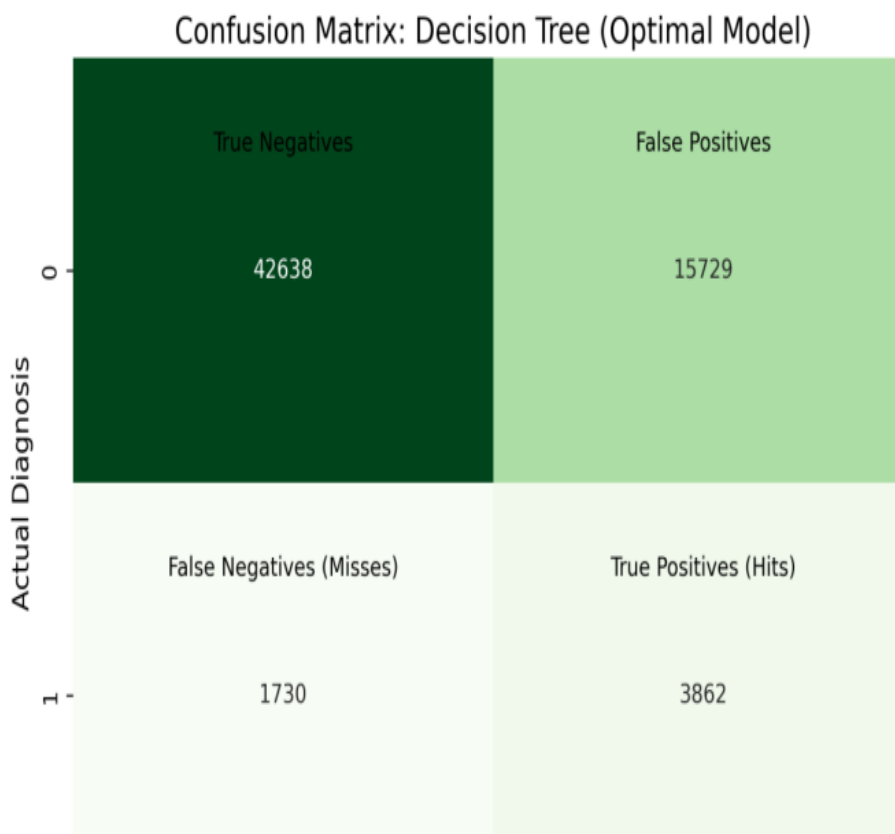


Figure 2: Confusion matrix for the Decision Tree model.

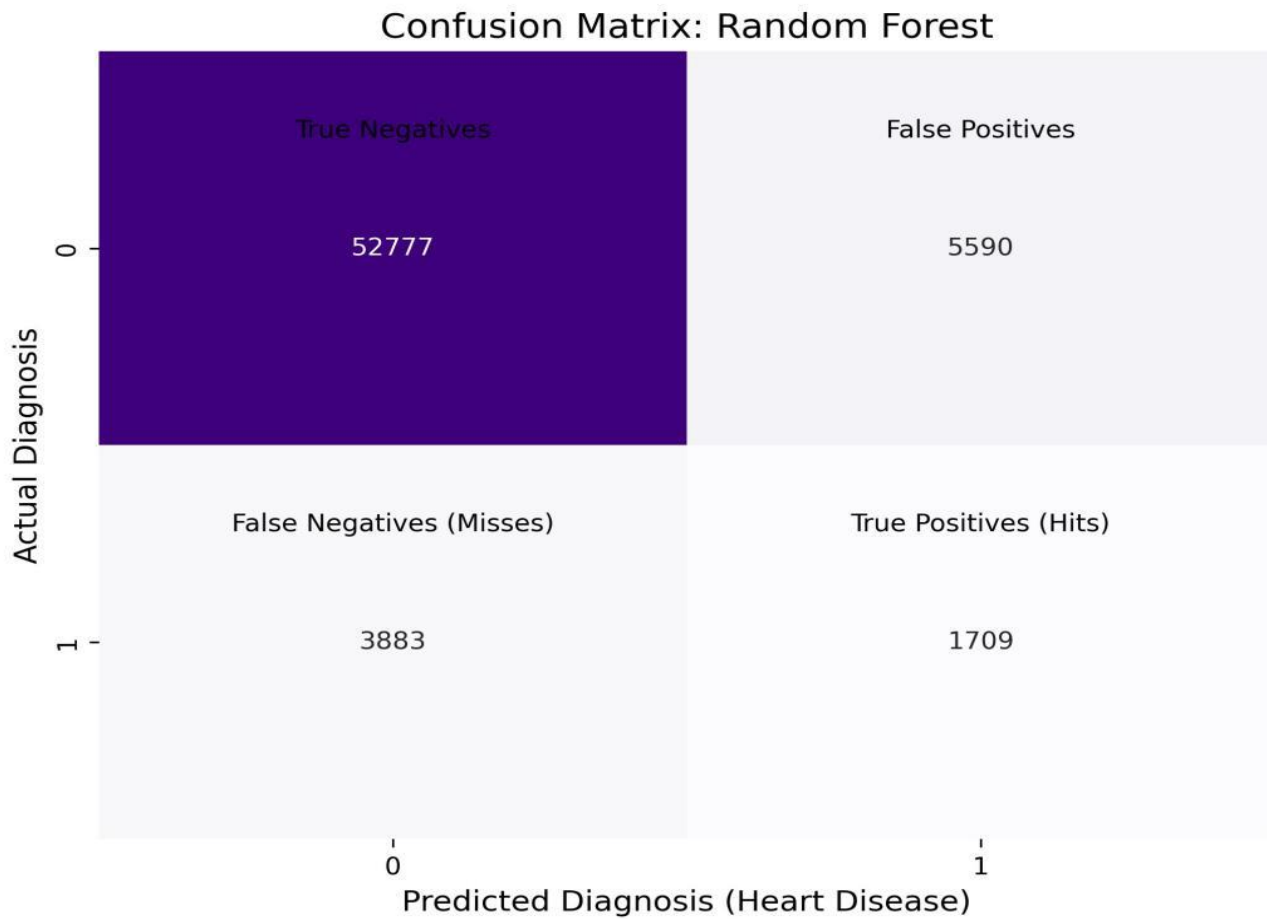


Figure 3: Confusion Matrix for the Random Forest.

The Decision Tree identifies 3,862 true positive cases and misses 1,730 positive cases. Logistic Regression identifies slightly more true positives (3,902) and has 1,690 false negatives, consistent with its marginally higher recall. However, Logistic Regression also generates 17,430 false positives. Random Forest produces far fewer false positives (5,590) but misses 3,883 positive cases, explaining its low recall despite its high accuracy. The matrices therefore reinforce the conclusion that the best model depends on the operational cost assigned to each type of error.

Table 3: Decision Tree confusion matrix before and after threshold optimization.

Metric	Threshold $\tau = 0.50$	Threshold $\tau = 0.36$
True Positive (TP)	3205	3862
False Negative (FN)	2387	1730
False Positive (FP)	11,245	15,729
True Negative (TN)	47,122	42,638
Recall(Sensitivity)	57.30%	69.06%
Specificity	80.74%	73.05%

Shifting the decision threshold from 0.50 to 0.36 substantially reduces the number of missed heart disease cases (false negatives) at the cost of increased false alarms. This trade off aligns with the clinical priority of minimizing false negatives consistent with the 10:1 cost ratio specified in the Bayesian analysis.

Cost-Sensitive Bayesian Optimization

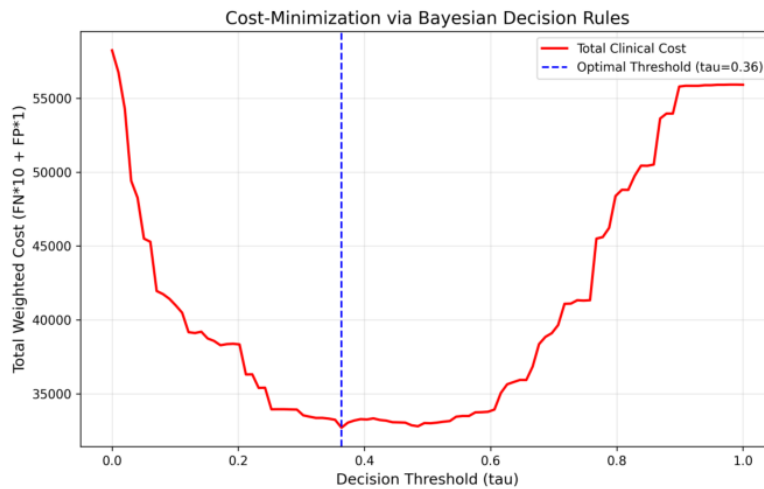


Figure 4: Cost-minimization curve for the Decision Tree model.

The cost-minimization curve shows the total weighted error cost across decision thresholds. [With false negatives penalized ten times more heavily than false positives, the minimum occurs at $\tau = 0.36$. Lowering the threshold from 0.50 to 0.36 makes the classifier more willing to flag potential cases, reducing the operational penalty associated with missed diagnoses. The result demonstrates how Bayesian decision analysis can convert a predictive model into a cost-aware decision-support rule.

Practical Implications and Limitations

The main practical implication is that accuracy should not be used as the sole criterion for selecting a heart disease screening model. Random Forest appears strongest under an accuracy-only criterion, but its low recall makes it less suitable where missed cases are costly. Logistic Regression provides high sensitivity but generates many false alarms. The Decision Tree provides the most balanced performance and offers a transparent sequence of decision rules that can be more readily examined by clinicians.

The analysis is based on a cross-sectional survey dataset, so the results should not be interpreted as establishing causality. The target variable is based on self-reported information, and the predictors do not represent a complete clinical diagnostic record. The 10:1 cost ratio is an operational assumption used for decision optimization rather than a universal clinical cost. External validation on independent populations and prospective clinical data would therefore be required before deployment.

CONCLUSION

This study developed an optimal machine-learning framework for early heart disease detection using the 2020 CDC BRFSS dataset. After preprocessing and SMOTE-based class balancing, Logistic Regression, Decision Tree and Random Forest models were compared using multiple performance criteria. Random Forest produced the highest accuracy (85.19%), but the Decision Tree produced the strongest balance of recall, F1-score and MCC, making it the preferred model for the stated screening objective. Bayesian cost-sensitive optimization further identified a threshold of $\tau = 0.36$ when missed diagnoses were assigned ten times the cost of false alarms. Overall, the findings show that interpretable modeling and decision-aware threshold selection can be more useful for healthcare decision support than optimizing accuracy alone.

REFERENCES

1. Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., Koohestani, A., Khozeimeh, F., Nahavandi, S., & Sarrafzadegan, N. (2019). A data base for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific Data*, 6, 227.
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
3. Centers for Disease Control and Prevention. (2020). Behavioral Risk Factor Surveillance System survey data. U.S. Department of Health and Human Services, CDC.
4. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
5. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6.
6. Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 973–978).
7. Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., Min, J. K., Tang, W. H. W., Halperin, J. L., & Narayan, S. M. (2019). Deep learning for cardiovascular medicine: A practical primer. *European Heart Journal*, 40(25), 2058–2073.
8. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
9. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
10. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574.
11. World Health Organization. (2017). Cardiovascular diseases. World Health Organization.