# Heart Disease Prediction Using Machine Learning Algorithms

**Sharanagouda N Patil[1*], Raju Hajere[2]**

[1]Department of Electronics and Communication Engineering, Govt. Womens Polytechnic, Kalaburagi, India. Email: snpgwpt@gmail.com
[2]Department of Electronics & TeleCommunication Engineering BMSIT&M, Bengaluru, India. Email: rajuhajare@bmsit.in

## Abstract

Cardiovascular disease ranks among the major causes of death in the global population, and timely and proper diagnosis is one of the most important healthcare goals. Machine learning has demonstrated good potential in helping clinicians to detect some latent trends in clinical data. This paper illustrates a heart disease predictive algorithm that applies the Logistic Regression (LR) and Random Forest (RF) algorithms. It uses a pipeline of data preprocessing, feature encoding, model training, cross-validation, and performance evaluation that is structured. Logistic Regression is interpretable and gives probabilistic results, whereas the random forest is more accurate by means of ensemble learning and non-linear modeling. The experimental analysis proves that the Random Forest tends to gain higher classification results, whereas the Logistic Regression can be still considered a reliable and explainable 9eline. The accuracy, precision, recall, F1-score, and ROC-AUC are also highlighted in the study as assessment measures that would be clinically relevant.

**Keywords**: Heart Disease Prediction, Logistic Regression, Random Forest, Machine Learning, ROC4AUC, Clinical Decision Support.

## 1. Introduction

Cardiovascular diseases (CVDs) have been identified to cause a considerable percentage of deaths worldwide annually, which is a great burden to healthcare systems [1]. The prevention and treatment of heart disease can be achieved at an early stage and the patient has a high chance of survival. Traditional risk assessment models like the Framingham Risk Score are based on the statistical regression model and a set cut off point [2]. Although they are effective in this case, they might not effectively capture complex interactions among various clinical variables.

Recent developments in machine learning have provided opportunities to utilize data-driven methods that can improve the predictive power by learning non-linear correlation on past patient data [3], [6]. Among them, Logistic Regression is still popular in clinical studies because it is simple and easy to interpret [4], but other methods that consistently perform better are ensemble methods or random forests who combine multiple decision trees into one single structure [5]. The paper compares and contrasts these two algorithms in prediction of heart disease based on structured clinical data.

The paper contributes to the area of machine learning-based prediction of cardiovascular diseases in three main ways. First, it introduces a useful and repeatable machine learning pipeline of heart disease prediction by the Logistic Regression (LR) and Random Forest (RF) algorithms. The

pipeline has a methodological rigor and reproducibility, as it systematically combines data preprocessing, feature encoding, training the model, cross-validation, and thorough performance evaluation.

Second, the paper gives a comparative analysis of LR and RF to predictive performance, interpretability, and probability calibration. Random Forest is also better at retaining non-linear associations as well as interactions among features whereas the Logistic Regression provides the chance to understand its results through transparent coefficients and predicts steady probabilities that are essential in clinical decision-making. Lastly, the paper focuses on reporting transparency and best practices of validation through alignment of model development and assessment with existing clinical prediction model guidelines. The recommendations are aimed at strong assessment, bias minimization, improving the reliability and practical implementation of machine learning models in a healthcare setting.

## 2. Related Work

The prediction of heart diseases with the help of machine learning has been the focus of numerous studies because structured clinical data is available and early, affordable diagnosis is required. A number of investigations have used publicly accessible benchmark datasets with the UCI Heart Disease repository being the most prominent one to test supervised learning algorithms in predicting cardiovascular risks [3]. This data has been adopted as a reference point of comparative analysis due to the inclusion of clinically relevancy attributes that include age, blood pressure, cholesterol level, type of chest pain and the electrocardiographic findings.

Logistic Regression (LR) is one of the classical methods that have found wide application in clinical studies because of its good statistical background, probability output and high interpretability [4], [16]. Medical decision-support systems have extensively utilized Logistic Regression models due to the ease with which the learned coefficients may be interpreted as the odds ratios and thus clinicians may comprehend the role of individual risk factors. A number of studies have found competitive heart disease classification with LR, especially in cases when the relationship between predictors and the outcome is informally linear [4]. As a result of these characteristics, LR is still used as a stable base model in healthcare analytics.

Conversely, by modeling non-linear relationships, as well as interactions among complex features, ensemble learning techniques and particularly RF have been shown to have a better predictive performance in numerous healthcare applications [5], [6]. Random Forest is constructed by training several decision trees using bootstrap samples, making it less prone to variance and more resilient than single-model methods. Earlier experiments on using RF in predicting heart diseases indicate higher accuracy in classification and generalization than the individual classifiers, especially when the dataset consists of diverse clinical features [5]. RF is also useful in medical datasets because of the capacity of the RF to deal with mixed feature types as well as the capability of RF to deal with noisy data.

The results of comparative studies that apply both Logistic Regression and Random Forest show that there is a definite trade-off between predictive power and interpretability. Random Forest tends to be more accurate and possess better ROC-AUC, whereas Logistic Regression has clearer decision

boundaries and predictable probability values, which are necessary to be applied in a clinic [6], [16]. Recent efforts have highlighted that, within the medical context, the use of models ought to select models based on performance/explainability ratios rather than on measures of accuracy alone.

In the research of predicting heart diseases, model evaluation and validation are still important challenges. In order to ensure that the generalization estimates are accurate, cross-validation techniques are usually used to determine the presence of bias in performance [8]. ROC is often used as a diagnostic performance measure, where ROC-AUC is used to determine the extent of discrimination at both classification thresholds and is appropriate when it comes to medical decision-making [9]. In addition, probability calibration has been utilized in clinical risk prediction since well-calibrated models offer more sound risk scores to stratify patients [10], [11].

This paper, on the one hand, is limited to the implementation of Logistic Regression and the comparison with Random Forest, as predictors of heart diseases as proposed by previous studies. The study presents a specialized analysis of the differences between statistical and ensemble learning paradigms as these two widely accepted algorithms are considered, yet the scope of the study remains narrow enough to comply with accepted standards of validation and reporting**.**

## 3. Dataset Description

The dataset used in this study is sourced from the UCI Heart Disease repository, a widely adopted benchmark dataset for cardiovascular disease prediction research [3]. The dataset is publicly available and compiled by the University of California, Irvine Machine Learning Repository. It consolidates clinical data collected from multiple research centers and has been used in numerous medical informatics and machine learning studies due to its structured format and clinically relevant features.

Each record in the dataset represents an individual patient and contains demographic, physiological, and diagnostic measurements that are commonly used in cardiovascular assessment. The target variable in the UCI Heart Disease dataset is a binary label indicating the presence or absence of heart disease, derived by consolidating original multi-level diagnosis outcomes into two clinically meaningful categories—a standard practice in predictive modeling studies [3], [4].

The dataset includes 14 attributes, comprising age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar status, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels (0–3) colored by fluoroscopy, and thalassemia status. These features are routinely collected in clinical settings and provide a comprehensive representation of risk factors associated with coronary heart disease.

Prior to model training, the dataset is inspected for missing or inconsistent values. Standard preprocessing techniques such as imputation for missing data, conversion of categorical variables via one-hot encoding, and feature scaling (for Logistic Regression only) are applied to prepare the data for analysis. Subsequently, the cleaned dataset is split into training and testing subsets using stratified sampling to preserve the original class proportions, which is essential for unbiased evaluation of performance metrics [8].

By using the UCI Heart Disease dataset, the study enables reliable comparison with existing work in the literature and ensures that results are reproducible and interpretable under well-accepted experimental conditions.

## 4. Methodology

This study adopts a structured machine learning methodology for heart disease prediction using Logistic Regression (LR) and Random Forest (RF) classifiers. The overall workflow consists of data preprocessing, feature transformation, model training, cross-validation, and performance evaluation. The methodological design follows established best practices for clinical prediction modeling to ensure reliability, reproducibility, and fair comparison between the implemented algorithms [8], [15].

### 4.1 Data Preprocessing

The raw clinical dataset is first examined for missing values, inconsistencies, and irrelevant entries. Missing numerical values are handled using statistical imputation techniques such as mean or median substitution, while categorical attributes are treated using mode imputation when required. Since the dataset contains both numerical and categorical features, categorical variables are converted into numerical form using one-hot encoding.Feature scaling is applied prior to Logistic Regression training to normalize the feature ranges and improve convergence during optimization. Standardization is performed by transforming features to zero mean and unit variance. Random Forest does not require feature scaling due to its tree-based architecture [4], [5].

### 4.2 Train–Test Split

The preprocessed dataset is divided into training and testing subsets using a stratified sampling strategy to preserve the original class distribution of heart disease and non-heart disease cases. This approach minimizes sampling bias and ensures that both classes are adequately represented in each subset. Typically, 70–80% of the data is used for training and the remaining portion for testing [8].

### 4.3 Logistic Regression Model

Logistic Regression is implemented as a baseline supervised learning model for binary classification. It estimates the probability of heart disease occurrence using a logistic function and learns model parameters through maximum likelihood estimation. To prevent overfitting and improve generalization, regularization techniques are employed during training. The probabilistic output of Logistic Regression enables direct interpretation of risk scores and supports clinical decision-making [4], [16].

### 4.4 Random Forest Model

Random Forest is implemented as an ensemble classifier consisting of multiple decision trees trained on bootstrap samples of the dataset. At each split, a random subset of features is selected to reduce correlation among trees and enhance model diversity. The final classification is obtained

through majority voting across all trees. This ensemble approach improves robustness and generalization performance, particularly in datasets with non-linear feature interactions and noisy measurements [5], [6].

### 4.5 Model Training and Cross-Validation

Both Logistic Regression and Random Forest models are trained using k-fold cross-validation on the training dataset. Cross-validation ensures stable performance estimation and reduces overfitting by evaluating models across multiple data partitions. Hyperparameters such as regularization strength for Logistic Regression and the number of trees and maximum depth for Random Forest are tuned during this stage [8].

### 4.6 Performance Evaluation

The trained models are evaluated on the independent test dataset using multiple performance metrics to ensure clinical relevance. These include accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC–AUC). ROC–AUC is used to assess the discriminative ability of the models across different classification thresholds, which is particularly important in medical diagnosis scenarios [9]. By employing standardized preprocessing, rigorous validation, and comprehensive evaluation, the proposed methodology enables a fair and meaningful comparison between Logistic Regression and Random Forest for heart disease prediction

## 5. Results and Discussion

This section presents the experimental results obtained from implementing Logistic Regression (LR) and Random Forest (RF) for heart disease prediction using the UCI Heart Disease dataset [3]. Model performance is evaluated using accuracy, precision, recall, F1-score, ROC–AUC, and average precision, which are widely accepted metrics for medical classification problems [9].

### 5.1 Quantitative Performance Analysis

The results show that Random Forest consistently outperforms Logistic Regression across all evaluation metrics. Random Forest achieves an accuracy of 88.59%**,** compared to 84.78% for Logistic Regression, indicating improved overall classification capability. Table I summarizes the classification performance of both models on the test dataset.

**Table I**
**Performance Comparison of Logistic Regression and Random Forest**

| Model | Accuracy | Precision | Recall | F1-score | ROC–AUC | Avg. Precision |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.8478 | 0.8426 | 0.8922 | 0.8667 | 0.9421 | 0.9226 |
| Random Forest | 0.8859 | 0.8716 | 0.9314 | 0.9005 | 0.9577 | 0.9631 |

The higher precision and recall values obtained by Random Forest demonstrate its superior ability to correctly identify both positive and negative heart disease cases.

### 5.2 Sensitivity and Clinical Importance.

Recall (sensitivity) is a very important measure in a clinical diagnosis because false negatives can cause missed diagnosis and delayed treatment. The Logistic Regression has a recall of 0.8922, which is good sensitivity; however, Recall by the use of Random Forest is even higher (0.9314), thus, it is more applicable in identifying the patients with risk of heart disease. This enhancement is an indication of the benefit of ensemble learning in modeling complicated associations amid risk elements of cardiovascular diseases [5], [6].

Random Forest (0.9005) is also superior to Logistic Regression (0.8667) in terms of F1-score, which is a balance score between precision and recall, which further supports the overall power of the ensemble model to address classification trade-offs.

### 5.3 Quality of Probability and Ability to Discriminate.

The ROC and AS values of the two models are quite high reflecting high performance in terms of discrimination. Logistic Regression has ROC -AUC of 0.9421 and Random Forest has an even higher value of 0.9577 indicating an enhanced differentiation of heart disease and other non-heart disease categories along varying thresholds [9].

Besides, the average precision score, which is a summary of the precision and recall performance, is better in the case of the Random Forest (0.9631) than in case of the Logistic Regression (0.9226). This finding indicates that the classification of high-risk patients given by the Random Forest gives a more credible ranking, which is particularly useful when it comes to clinical screening and triage [10], [11].

### 5.4 Interpretability/performance Trade-off.

Although the predictive ability of the Random Forest is better, the Logestic Regression is more predictable, and this feature is a critical need in most medical facilities. The odds ratios represented by the coefficients of Logistic Regression are easily interpreted, which allows clinicians to comprehend the impact of each risk factor on the prediction of heart diseases [4], [16]. Random Forest is more precise and less transparent with decision boundaries yet it frequently needs extra explain ability methods to be clinically interpreted.

In general, it can be observed that the findings exhibit a distinct trade-off between predictive performance and interpretability. Random Forest has increased accuracy, recall, ROC–AUC, and average precision, therefore, it is applicable in automated heart disease risk predictive systems. Although a little less accurate, Logistic Regression provides good results with an extra benefit of transparency and easy interpretation. It should therefore be implemented that the choice of a suitable model must be made depending on the application context either with the priority of maximum predictive accuracy or interpretability and clinical trust.

### 5.5 Confusion Matrix and ROC

The visual representation of Fig. 1a and b confusion matrix of Logistic Regression represents the distribution of true positives, true negatives, false positives, and false negatives. The model is highly recalled (0.8922) and thus successful in identifying heart disease cases, but has balanced precision. (0.8426). The presence of some false negatives suggests that linear decision boundaries may miss subtle non-linear patterns in the data.

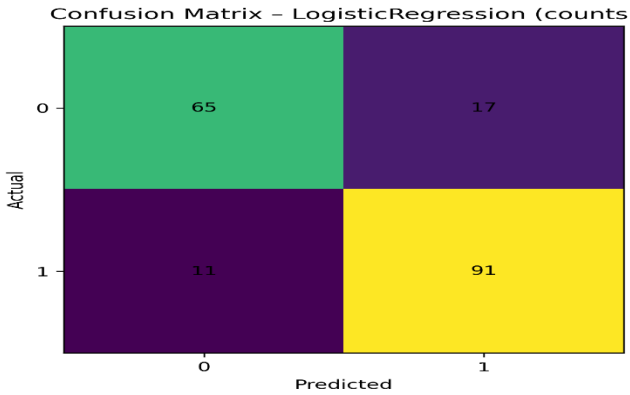


Figure 1a: Confusion Matrix Logistic Regression

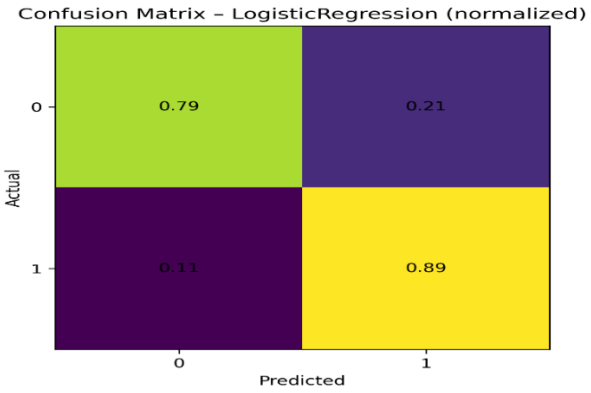


Figure 1b: Confusion Matrix Logistic Regression (Normalized)

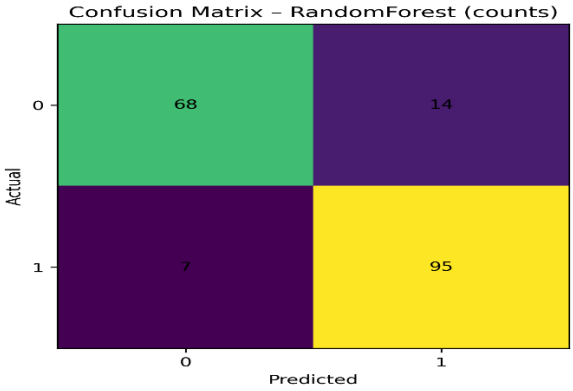


Figure 2a: Confusion Matrix Random Forest

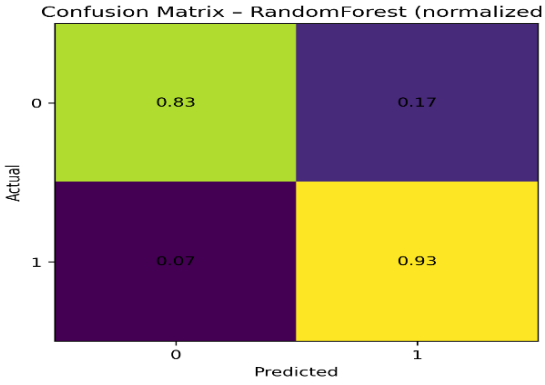


Figure 2b: Confusion Matrix Random Forest (Normalized)

This Fig. 2a and b shows the confusion matrix for the Random Forest classifier. Compared to Logistic Regression, Random Forest produces fewer false negatives and false positives, consistent with its higher accuracy (0.8859) and recall (0.9314). The improved classification performance reflects the ensemble's ability to capture complex interactions among clinical features, leading to more reliable predictions.

The Fig. 3a and b illustrates the Receiver Operating Characteristic **(**ROC) curves for Logistic Regression and Random Forest. The ROC curve plots the true positive rate (recall) against the false positive rate at different decision thresholds, providing a threshold-independent measure of

discrimination. Both models demonstrate strong discriminative ability, as indicated by ROC–AUC values above 0.94. Random Forest achieves a higher ROC–AUC (0.9577) compared to Logistic Regression (0.9421), indicating superior class separation and robustness across thresholds.

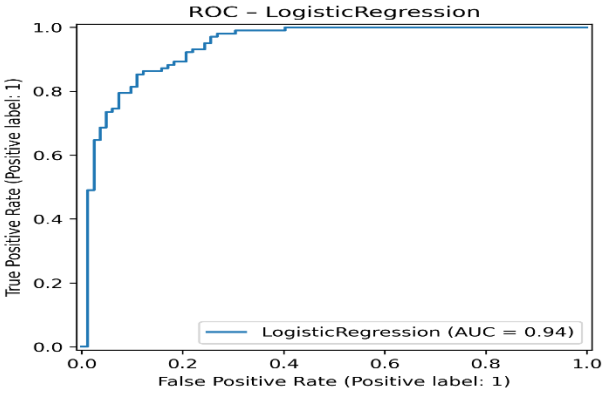


Figure 3a: ROC of Logistic Regression

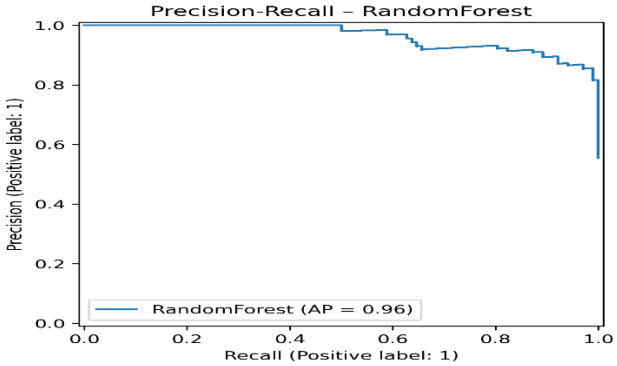


Figure 3b: ROC of Logistic Regression (Normalized)

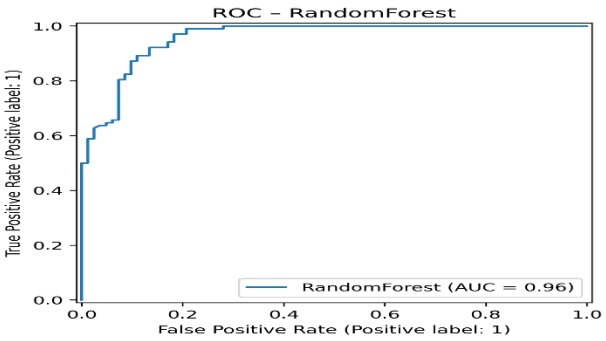


Figure 4a: ROC of Random Forest

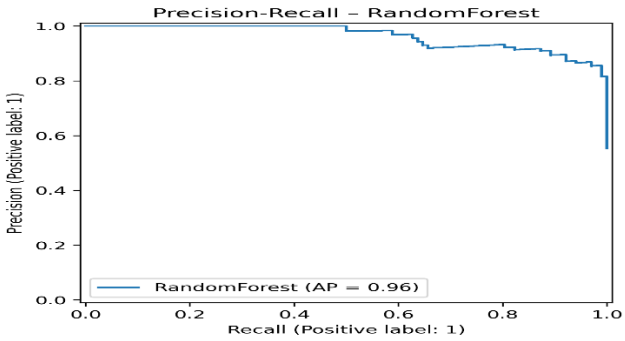


Figure 4b: ROC of Random Forest (Normalized)

The Precision-Recall (PR) curves included in the Fig. 4a and b are especially informative in medical diagnosing when the distribution of classes can be unequal. The curve illustrates the trade-off between precision and recall of varying thresholds. Random Forest has a greater average precision (0.9631) as compared to Logistic Regression (0.9226), which proves to be more reliable in identifying those patients that are at high risk as well as having lower false positives. This shows that Random Forest is appropriate in screening and triage application.

## 6. Conclusion

The current paper introduced a machine learning-based predictive method of heart disease based on the Logistic Regression and Random Forest algorithms applied on the UCI Heart Disease data. Reliable and reproducible results were done by systematic pipeline, which comprised data preprocessing, stratified training, testing, cross-validation, and multi-metric evaluation. The results of the experiment prove that both models perform well in terms of predictive power; nevertheless, the Random Forest always wins in all the measures of evaluation, being more accurate, recalls, and

F1-score, having higher ROC-AUC, and average precision. The high accuracy of Random Forest demonstrates the efficiency of ensemble learning to determine the non-linear relationships and complicated interactions among cardiovascular risk factors.

Although Logistic Regression is less accurate than other models, it is still a useful model since it is interpretable and has stable probabilities, which are critical in clinical decision-making and risk assessment that can be explained. The findings thus highlight a trade-off between predictive performance and transparency that is practical and implies that the choice of a model should be informed by the desired clinical use. Random Forest would be more suitable in terms of sensitivity and discriminative power to use as an automated screening system, but Logistic Regression would be better to use in situations where an interpretable decision-support tool is needed.

The further research will target external validation with bigger and more diverse clinical data, integration of enhanced calibration and explain ability methods, and testing of model equity among demographic subgroups. The extensions will also promote the reliability, generalizability, and clinical applicability of the machine learning-based heart disease prediction systems.

## References

[1] [1] World Health Organization, “Cardiovascular diseases (CVDs),” WHO, 2021.

[2] R. B. D’Agostino Sr. *et al*., “General cardiovascular risk profile for use in primary care: The Framingham Heart Study,” *Circulation*, vol. 117, no. 6, pp. 743–753, Feb. 2008,

[3] UCI Machine Learning Repository, “Heart Disease dataset,” Univ. of California, Irvine. UCI Machine Learning Repository

[4] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[5] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[7] F. Pedregosa *et al*., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[8] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. IJCAI*, 1995.

[9] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[10] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. ICML*, 2005.

[11] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[14] R. F. Wolff *et al*., "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019.

[15] K. G. M. Moons *et al*., "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration," *Ann. Intern. Med.*, 2015, doi: 10.7326/M14-0698.

[16] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-16399-0.

[17] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, 2010.