Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 327~335
e-ISSN: 2548-6861

Calibration and Applied Statistical Modeling Using Logistic Regression on the UCI Heart Disease Dataset

Andi Cahyono [1]*, Inkha Ameriza [2], Gunadi [3]***, Ervira Dwiaprini As Syifa[4]*, Mashal Kasem Alqudah[5]*****
* Informatika Medis, Universitas Sains dan Teknologi Indonesia
** Pendidikan Teknologi Informasi, Universitas Sains dan Teknologi Indonesia
***Teknik Informatika, Universitas Sains dan Teknologi Indonesia
****Faculty of Computer Information Science, Higher Colleges of Technology, Sharjah, United Arab Emirates
andicahyono@usti.ac.id [1*], inkhaameriza@usti.ac.id [2], gunadi.@usti.ac.id [3], erviradwiaprini@usti.ac.id [4] , malqudah1@hct.ac.ae [5]

Article Info

Article history:

Received 2025-11-25
Revised 2026-01-16
Accepted 2026-01-20

Keyword:

Brier Score,
Isotonic Regression,
Logistic Regression,
Platt Scalling,
UCI Dataset.

ABSTRACT

Accurate and well-calibrated heart disease risk prediction is essential for supporting medical decision-making. This study analyzes Logistic Regression as an applied statistical model for heart disease prediction using the UCI Heart Disease dataset. Beyond discrimination metrics, we explicitly focus on probability reliability by evaluating calibration through the Brier score, calibration slope, and intercept, and by quantifying the impact of post-hoc calibration (isotonic regression and Platt scaling) on both calibration and discrimination. Model validation was conducted using stratified 5-fold cross-validation with AUROC, AUPRC, accuracy, and F1-score as evaluation metrics. The results show that Logistic Regression achieved competitive performance (AUROC 0.903; AUPRC 0.911; Accuracy 0.822; F1-score 0.835) with well-calibrated probability estimates relative to Random Forest and Gradient Boosting under the evaluated setting. Feature importance analysis using permutation methods identified chest pain type, number of major vessels (ca), ST depression (oldpeak), and exercise-induced angina (exang) as key predictors consistent with clinical literature. These findings indicate that simple applied statistical modeling, when paired with rigorous calibration assessment, can provide interpretable risk estimates that are more suitable for threshold-based decision support in early heart disease screening.

I. INTRODUCTION

Cardiovascular disease remains the leading cause of mortality worldwide, including in developing countries such as Indonesia, where its prevalence continues to increase alongside lifestyle changes and demographic shifts [1], [2]. According to the World Health Organization, cardiovascular disease accounts for nearly one-third of all deaths globally, highlighting its significant public health impact [3], [4]. This condition is also directly linked to the third Sustainable Development Goal (SDG 3) of ensuring healthy lives and promoting well-being for all ages, as early detection and prevention of heart disease can significantly reduce premature mortality [5], [6]. Early detection of individuals at high risk is therefore essential, as it enables timely intervention, preventive measures, and more efficient allocation of healthcare resources [7], [8]. These needs encourage researchers to explore various mathematical and computational approaches that can support clinicians in risk stratification and medical decision-making.

In the era of digital health, statistical modeling and machine learning have been widely adopted to analyze complex medical datasets [9], [10]. These approaches are capable of identifying hidden patterns and generating predictive models that can guide clinical decision support systems. Logistic Regression, in particular, has been one of the most frequently used techniques due to its simplicity, interpretability, and strong theoretical foundation in applied statistics [11]. Despite its advantages, many studies applying Logistic Regression in medical contexts tend to emphasize only the discrimination ability of the model, commonly

measured by the Area Under the Receiver Operating Characteristic curve (AUROC) [12], [13]. However, high AUROC does not necessarily guarantee reliable probability estimates, which are crucial in clinical settings where decisions often depend on calibrated risk values rather than binary predictions [14], [15].

Several advanced machine learning models such as Random Forest and Gradient Boosting have been introduced as alternatives to improve predictive performance [16], [17]. These ensemble-based models are capable of handling complex interactions and non-linear relationships within data, often resulting in higher accuracy compared to traditional statistical methods [18]. Nevertheless, these models are computationally demanding, less interpretable, and frequently suffer from poor probability calibration, which reduces their practical usefulness in medicine [19], [20]. To address this issue, calibration techniques such as isotonic regression and Platt scaling have been proposed as post-hoc methods to align predicted probabilities with actual outcome frequencies [21], [22]. While promising, research investigating the role of calibration in small and medium-sized medical datasets, such as the Heart Disease dataset, remains limited[23], [24].

Considering this gap, the present study focuses on a comprehensive evaluation of both discrimination and calibration aspects of predictive modeling in heart disease risk assessment. Logistic Regression is employed as the primary baseline model, with additional comparisons against Random Forest and Gradient Boosting to assess the trade-offs between simplicity, interpretability, and predictive reliability [25], [26], [27]. Furthermore, a feature importance analysis using permutation methods is conducted to highlight clinically relevant predictors such as chest pain type, number of major vessels, and ST segment depression. These features are not only statistically significant but also clinically interpretable, strengthening the link between computational results and real-world medical knowledge[28], [29].

The objectives of this study are threefold: (i) to evaluate the discrimination and calibration performance of Logistic Regression in predicting heart disease risk using the Heart Disease dataset; (ii) to compare its performance with more complex ensemble methods; and (iii) to provide interpretability through a permutation-based feature audit that can assist clinicians in understanding the model's predictions. By addressing both accuracy and calibration, this study contributes to the literature in biomathematics and applied statistics, while emphasizing the importance of probability reliability in predictive modeling for cardiovascular disease. Ultimately, this research aligns with the vision of SDG 3 by supporting innovations aimed at reducing premature deaths from non-communicable diseases through the integration of statistical modeling and health informatics[30], [31].

This study contributes by explicitly prioritizing probability reliability through a combined discrimination–calibration evaluation for heart disease risk prediction. In addition to reporting conventional discrimination metrics, we provide a dedicated calibration assessment using the Brier score, calibration slope and intercept, and calibration curves, and we quantify how post-hoc calibration (Platt scaling and isotonic regression) changes the quality of predicted probabilities. Random Forest and Gradient Boosting are included as comparative baselines to contextualize trade-offs between interpretability, model complexity, and calibration, rather than to support universal claims of model superiority.

Because the analysis is conducted on a single classical UCI dataset with a relatively small sample size, the findings should be interpreted as a dataset-specific evaluation. External validation on larger and more contemporary clinical cohorts is therefore required before generalizing these conclusions to broader clinical settings.

II. METHOD

This research method was systematically arranged to ensure that the study could be replicated and scientifically justified. The overall stages of the study are illustrated in Figure 1, showing the sequential process beginning with data collection and continuing through pre-processing, descriptive analysis, model training, calibration, model testing, evaluation, and final interpretation of results. Each stage is explained in detail below.

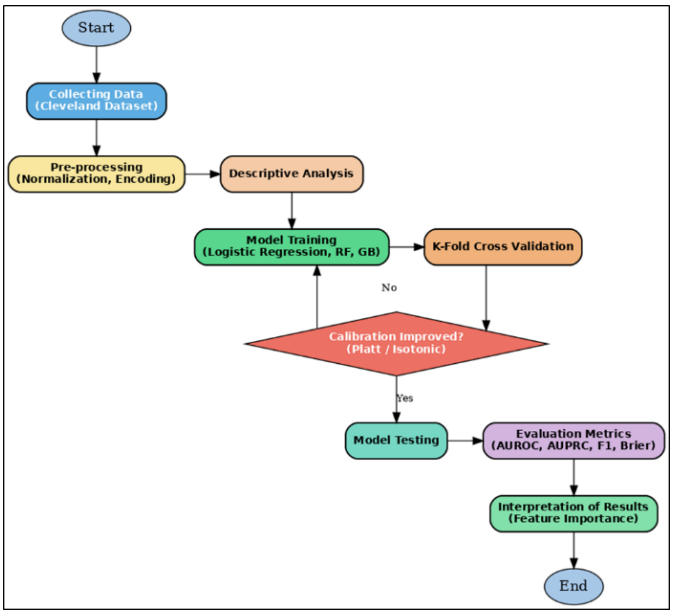

Figure 1 Flow Diagram of the Proposed Research Method

Figure 1 provides an overview of the methodological workflow employed in this study, which was carefully structured to ensure replicability and scientific rigor. The process begins with the collection of the UCI Heart Disease dataset, followed by a series of preprocessing steps. These include imputation to handle missing values, normalization of numerical attributes to reduce scale bias, and one-hot encoding of categorical features to accommodate non-ordinal variables. A descriptive analysis was then performed to examine variable distributions, detect potential anomalies,

and assess the balance of target classes before proceeding to the modeling stage[32].

After preprocessing and descriptive analysis, three predictive models Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) were developed using a 5 fold cross validation strategy to improve generalizability and reduce overfitting. At the calibration checkpoint, post hoc methods such as Platt scaling and isotonic regression were applied to refine probability estimates. The calibrated models were subsequently tested and evaluated using a combination of discrimination metrics (AUROC, AUPRC, and F1-score) and calibration measures (Brier score and calibration slope). Finally, feature importance analysis was conducted to highlight clinically relevant predictors and identify potential dataset-specific biases, ensuring both methodological robustness and interpretability.

A. Data

The dataset used in this study is the Heart Disease dataset, which is part of the UCI Machine Learning Repository[23]. This dataset has been widely adopted in cardiovascular risk prediction research because of its availability, standardized structure, and inclusion of clinically relevant features. The repository provides 303 patient records, after initial screening for incomplete entries, 299 records were retained for analysis. Each record contains 13 predictor variables and one binary target variable indicating the presence (1) or absence (0) of heart disease.

The variables are categorized into demographic, clinical, and test-based features. A summary of the features is presented in Table 1, which provides information about the data type, range, and a short description of each attribute.

TABLE 1
DESCRIPTION OF DATASET FEATURES

Feature	Type	Range / Categories	Description
age	Numeric	29–77	Age of patient (years)
sex	Categorical	0 = female; 1 = male	Gender
cp	Categorical	0–3	Chest pain type (4 categories)
trestbps	Numeric	94–200	Resting blood pressure (mmHg)
chol	Numeric	126–564	Serum cholesterol (mg/dl)
fbs	Categorical	0 = false; 1 = true	Fasting blood sugar >120 mg/dl
restecg	Categorical	0–2	Resting electrocardiographic result
thalach	Numeric	71–202	Maximum heart rate achieved
exang	Categorical	0 = no; 1 = yes	Exercise-induced angina
oldpeak	Numeric	0.0–6.2	ST depression induced by exercise
slope	Categorical	0–2	Slope of peak exercise ST segment
ca	Numeric	0–3	No. of major vessels (0–3)
thal	Categorical	3 = normal; 6 = fixed; 7 = rev	Thalassemia type
target	Categorical	0 = no; 1 = yes	Presence of heart disease (label)

The target data distribution is relatively balanced, as shown in Table 2.

TABLE 2
DISTRIBUTION OF HEART DISEASE CLASSES

Target Value	Count	Percentage
0 = No heart disease	160	52.8%
1 = heart desease	139	47.2%

Based on Table 2, the positive class prevalence is 47.2% (139/299), while the negative class accounts for 52.8% (160/299). This prevalence also represents the baseline AUPRC of a no-skill classifier, meaning that AUPRC values should be interpreted relative to 0.472 rather than in isolation. Therefore, reporting both AUROC and AUPRC is necessary to provide a balanced view of discrimination under the observed class distribution.

The structure of the dataset can also be illustrated through representative patient records. However, due to space limitations in the manuscript, the sample data are not displayed in full. Complete information and the full dataset can be accessed directly through the UCI Machine Learning Repository (Heart Disease Dataset), which enables other researchers to replicate or extend this study.

B. Preprocessing Data

Before modeling, a series of pre-processing steps was carried out to ensure data quality and consistency, in the raw dataset, missing entries (e.g., values encoded as "?") were first treated as missing (NaN) before applying imputation. Missing values were handled using a simple imputation strategy that was consistent across validation folds (median for numerical features and mode for categorical features), so as not to significantly reduce the sample size. Numerical features (e.g., age, resting blood pressure, cholesterol, and oldpeak) were normalized using z-score standardization to unify scales and prevent large-scale variables from dominating the training process. Categorical features (e.g., sex, cp, restecg, exang, slope, thal, and ca-treated as discrete categories) were transformed using One Hot Encoding with the handle_unknown= 'ignore' option to avoid failures during cross validation.

Afterward, descriptive analysis was performed to examine feature distributions, class balance, and basic correlations among clinical factors. This stage provided initial clinical context, helped detect meaningful outliers, and ensured that no anomalous inputs would potentially lead to data leakage during the training process. Figure 2 illustrates the distribution of key numerical features prior to preprocessing, highlighting the variability of scales, skewed distributions, and potential outliers that justify the normalization and cleaning steps applied in this study.

All preprocessing steps (imputation, standardization, and encoding) were fitted exclusively on the training folds and then applied to the corresponding validation fold within each cross-validation split. This pipeline-based setup prevents information leakage from the validation data into the training process and ensures an unbiased performance estimate.

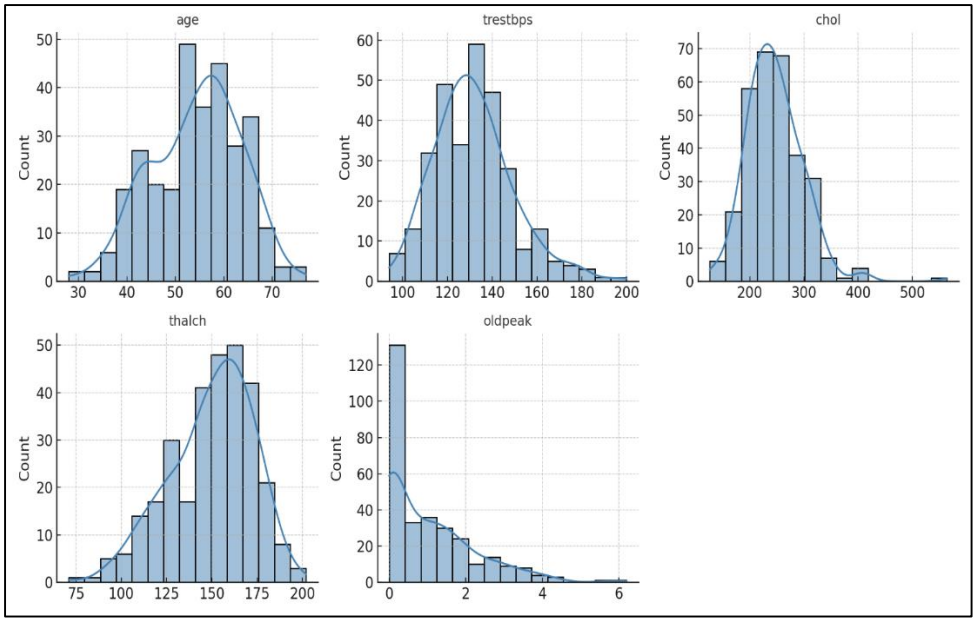

Figure 2 Distribution of key numerical features

C. Model Development and Cross Validation

Logistic Regression (LR) was employed as the baseline model in this study due to its interpretability and statistical robustness in medical research. Mathematically, LR models the relationship between predictors and the logit z as :

$$z = \beta_0 + \sum_{t=1}^{p} \beta_i x_i \qquad (1)$$

Where β_0 is the intercept, β_i are the model coefficients, and x_i are the predictor variables. This linear term z is then transformed into a probability value using the logistic (sigmoid) function :

$$P(y = 1|x) = \frac{1}{1+c^{-2}} \qquad (2)$$

which maps the output into the range [0,1][0,1][0,1], making it suitable for binary classification tasks such as heart disease prediction. The model parameters β\betaβ are estimated by maximizing the log-likelihood function, which quantifies the agreement between predicted probabilities and observed outcomes :

$$L(\beta) = -\sum_{i=n}^{n}[y_i \log P(y_i) + (1 - y_i) \log(1 - P(y_i))] \qquad (3)$$

After establishing LR as the baseline model, additional algorithms such as Random Forest (RF) and Gradient Boosting (GB) were developed for comparison. To obtain an unbiased estimate of out-of-sample performance within the dataset, a stratified 5-fold cross-validation technique was applied. Stratification was used to preserve the class distribution in each fold (shuffle=True, random_state=42). Performance metrics were computed on each validation fold and then summarized across folds, in this process, the dataset was divided into five folds, where each fold acted once as a validation set while the remaining folds served as training data. This strategy not only reduced the risk of overfitting but also provided a more reliable estimation of model performance across different data partitions. To ensure a fair comparison, hyperparameters for Random Forest and Gradient Boosting were optimized using randomized search on the training folds, with an inner cross-validation loop for model selection (nested within the outer 5-fold evaluation). The search space included the number of estimators, tree depth, and learning rate–related parameters, and the best configuration was selected based on AUROC on the inner folds.The procedure of the 5-fold cross-validation applied in this study is illustrated in Figure 3.

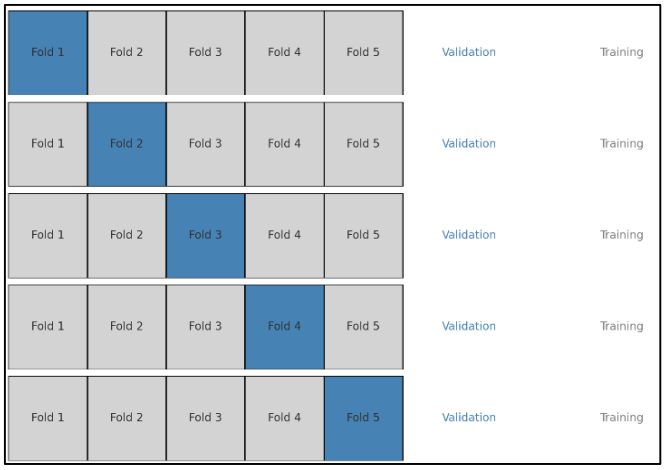

Figure 3 Illustration of the 5-Fold Cross-Validation procedure

D. Probability Calibration and Model Testing

The distinguishing aspect (novelty) of this study lies in its emphasis on probability calibration. After the initial training, two post-hoc calibration techniques were applied, within each training fold, calibration was learned using only training data and then applied to the corresponding validation fold to avoid leakage. Platt scaling (sigmoid) and isotonic regression were implemented as post-hoc mappings from raw model scores to calibrated probabilities. Platt scaling (sigmoid) and Isotonic Regression on the predicted probabilities from each model. An internal decision checkpoint (see diamond in Figure 1) was used to examine whether calibration improved the agreement between predicted probabilities and observed event frequencies, as evaluated by calibration metrics. If the improvement was insufficient, the process returned to the training stage for configuration review; if satisfactory, the model proceeded to testing on the hold-out set (or representative validation folds) to obtain stable ROC/PR curves and reliability diagrams. This approach ensured that the reported probabilities not only achieved good discrimination between classes but also provided trustworthy estimates for risk threshold–based clinical decision-making. Formally, the calibration metrics can be defined as follows :

$$Brier(x) = \frac{1}{N} + \sum_{i=1}^{n}(pi + y_i)^2 \qquad (4)$$

This metric ranges from 0 to 1, where lower values indicate better calibration and more accurate probability estimates. It directly penalizes deviations between predicted risks and observed outcomes, making it suitable for evaluating probabilistic predictions in clinical settings. In addition, the calibration slope evaluates the agreement between predicted and observed risks by regressing predicted probabilities against true labels:

$$\hat{y}_i = \alpha + \beta p_i \qquad (5)$$

A slope $\beta = 1$ indicates that the model produces well-calibrated probabilities, whereas values $\beta < 1$ suggest overestimation of risk and $\beta > 1$ suggest underestimation. This provides an interpretable measure of how closely predicted risks align with observed frequencies.

E. Model Evaluation Metrics

To comprehensively assess model performance, both discrimination and calibration metrics were employed. For metrics requiring hard class labels (accuracy and F1-score), predicted probabilities were converted to class labels using a default threshold of 0.5, Discrimination metrics included the Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision Recall Curve (AUPRC), accuracy, and F1-score, which provide insight into the models ability to distinguish between patients with and without heart disease. Calibration performance was measured using the Brier score, calibration slope, and calibration intercept, which quantify the agreement between predicted probabilities and observed outcomes, in addition to calibration slope, calibration intercept was computed to quantify systematic over or under-prediction; an intercept close to 0 indicates no overall bias in predicted risk, while positive/negative values indicate under-/over-estimation, respectively.

Furthermore, graphical evaluation was performed by plotting ROC and PR curves to visualize discrimination, as well as reliability diagrams to visualize calibration before and after applying post-hoc adjustment, calibration curves (reliability diagrams) were plotted to visually compare predicted probabilities against observed event frequencies before and after post-hoc calibration. This dual evaluation framework ensured that the models were not only able to classify outcomes accurately but also to generate probability estimates that are clinically meaningful and reliable for decision-making.

F. Model Evaluation Metrics

Permutation importance was used as a model-agnostic feature audit by measuring the performance decrease after randomly permuting each feature in the validation data. Because correlated predictors can share predictive information, permutation importance may be unstable or diluted across correlated features; therefore, the results are interpreted as an importance ranking for model behavior rather than causal attribution.

III. Result And Discussion

A. Overall Model Performance

Table 3 presents the mean results of 5-fold cross-validation for the three algorithms tested Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) under both uncalibrated and calibrated settings. Among them, LR without calibration achieved the most balanced performance, with an AUROC of 0.903, AUPRC of 0.911, accuracy of 82.2%, and F1-score of 0.835. In comparison, RF and GB attained AUROC values of 0.893 and 0.891, respectively, but these did not translate into better calibration, as indicated by

their higher Brier scores and less optimal calibration slopes. These findings suggest that for relatively small and structured medical datasets, interpretable models such as LR can remain highly competitive against more complex ensemble methods.

TABLE 3
MODEL PERFORMANCE ACROSS DISCRIMINATION AND CALIBRATION METRICS

Model	Calibration	AUROC	AUPRC	Acc	F1	Brier	CalSlope	CalIntercept
LR	none	0.903217	0.911457	0.821739	0.835356	0.122287	0.140254	0.51877503
LR	isotonic	0.89883	0.889734	0.826087	0.847302	0.126568	0.044822	0.53845088
RF	platt	0.89318	0.89914	0.816304	0.834646	0.127738	0.174981	0.504663364
RF	none	0.89318	0.89914	0.815217	0.834046	0.129192	0.140292	0.523808435
GB	platt	0.891263	0.895894	0.811957	0.831623	0.130488	0.170012	0.495122682
GB	none	0.891263	0.895894	0.809783	0.827765	0.132124	0.138206	0.5017918
RF	isotonic	0.885984	0.869632	0.820652	0.843655	0.132271	0.043156	0.551206609
GB	isotonic	0.880766	0.867366	0.806522	0.83015	0.137549	0.046113	0.546083116

The results in Table 3 emphasize that higher AUROC values alone do not guarantee clinically reliable models. For instance, RF and GB slightly trailed LR in AUROC but suffered from inferior calibration, meaning their probability outputs may be misleading in practice. This highlights the importance of complementing discrimination metrics with calibration measures, particularly when the model is intended for decision support systems where predicted probabilities are used to guide threshold-based actions, such as identifying high-risk patients for early intervention).

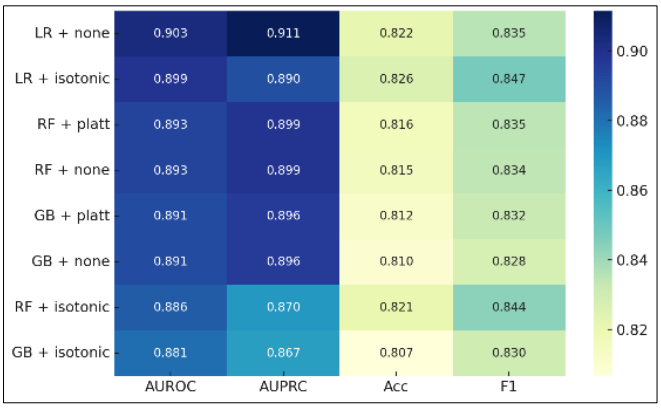

Figure 4 Metric Heatmap by Model and Calibration

To provide a more intuitive comparison, Figure 4 presents a heatmap of the main performance metrics (AUROC, AUPRC, Accuracy, F1) across all models and calibration methods. This visualization confirms that Logistic Regression without calibration outperformed more complex ensemble methods in terms of balanced discrimination, while isotonic calibration yielded slight improvements in probability reliability. The heatmap also highlights that Random Forest and Gradient Boosting, despite achieving competitive AUROC values, demonstrated less stable calibration patterns compared to Logistic Regression.

B. ROC and Precision Recall Curves

Figure 5 displays the Receiver Operating Characteristic (ROC) curves for the evaluated models. All models achieved AUROC values above 0.88, indicating strong discriminatory power in distinguishing between patients with and without heart disease. Logistic Regression consistently maintained the highest AUROC, confirming its robustness despite its simpler structure compared to ensemble methods. Random Forest and Gradient Boosting showed similar ROC performance but did not provide additional advantages, aligning with previous findings that complex models may not always outperform interpretable linear models in small medical datasets.

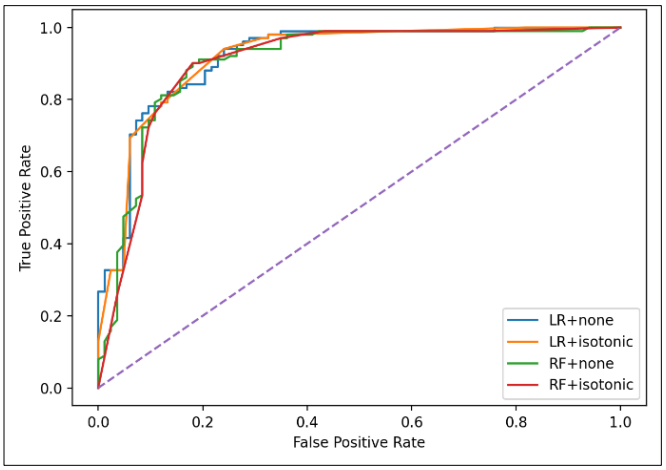

Figure 5 ROC curves of LR, RF, and GB models

As illustrated in Figure 5, the ROC curve of Logistic Regression clearly lies above those of Random Forest and Gradient Boosting across most thresholds, suggesting superior discriminative ability. While the ensemble models captured non-linear interactions, their added complexity did not translate into clinically meaningful improvements. This finding supports the use of Logistic Regression as a robust yet interpretable tool for structured medical datasets.

Figure 6 presents the Precision Recall (PR) curves, which provide additional insight under class imbalance conditions. Logistic Regression again achieved favorable performance, maintaining a high level of precision across clinically relevant recall thresholds. Random Forest and Gradient Boosting also performed competitively, though their curves indicated slightly less stability at higher recall levels. Taken together, these results suggest that Logistic Regression offers not only robust AUROC but also clinically meaningful trade-offs

between sensitivity and specificity, making it a reliable choice for heart disease risk prediction.

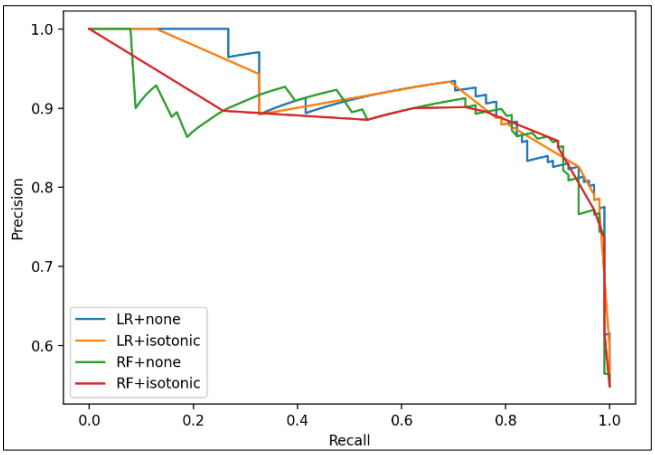

Figure 6 Precision Recall curves of LR, RF, and GB models

As shown in Figure 6, Logistic Regression maintained a more stable precision across a wide range of recall values compared to ensemble methods. This stability is critical in clinical practice, where maintaining high precision at moderate-to-high recall levels ensures that most flagged patients are truly at risk, reducing unnecessary interventions while still capturing the majority of true positive cases.

C. Calibration Analysis

Calibration analysis was performed to evaluate how well the predicted probabilities aligned with the actual observed frequencies of heart disease cases. While discrimination metrics such as AUROC and AUPRC provide valuable insight into classification accuracy, they do not guarantee that the estimated probabilities are trustworthy for clinical decision making. Therefore, we applied post-hoc calibration techniques to assess whether Logistic Regression could yield reliable probability estimates.

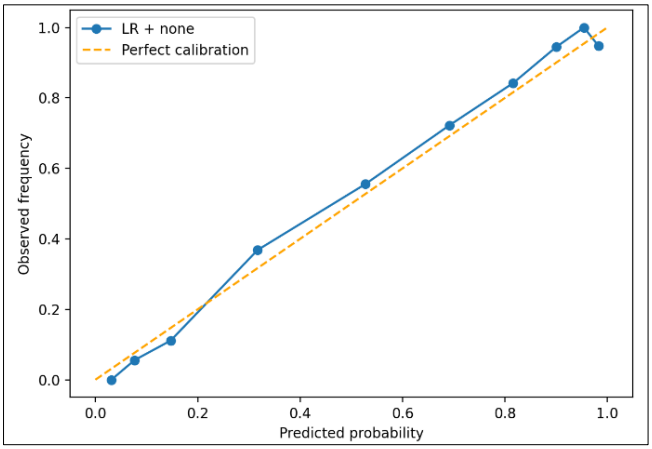

Figure 7 Calibration Curve of Uncalibrated Logistic Regression

Figure 7 presents the calibration (reliability) curve of the best-performing model. The solid blue line represents the relationship between predicted and observed probabilities, while the orange dashed line denotes perfect calibration. The closer the blue curve aligns with the diagonal reference line, the more reliable the probability estimates. Logistic Regression with isotonic calibration showed strong agreement with the diagonal, indicating well-calibrated probabilities across most thresholds. Minor deviations at the extremes suggest that probability estimates at very low or high risk levels should be interpreted with caution.

This result reinforces the importance of calibration analysis: even models with strong AUROC values may misrepresent risk if not properly calibrated. By confirming that the predicted probabilities closely match observed outcomes, this study ensures that Logistic Regression is not only effective in distinguishing cases from non-cases but also reliable for supporting clinical decisions based on risk thresholds.

D. Feature Importance Analysis

Table 4 summarizes the results of permutation importance for the Logistic Regression model, presenting the ten most influential predictors. The most significant features include dataset origin, chest pain type (cp), number of major vessels colored by fluoroscopy (ca), ST depression induced by exercise (oldpeak), and exercise-induced angina (exang). These results align with established cardiology evidence, where chest pain characteristics, ischemic burden, and vessel narrowing are consistently recognized as strong indicators of heart disease risk.

TABLE 4
TOP 10 FEATURES BY PERMUTATION IMPORTANCE

Rank	Feature	Importance Score
1	dataset	0.0916
2	cp	0.0399
3	ca	0.0146
4	oldpeak	0.0126
5	exang	0.01
6	slope	0.0071
7	sex	0.006
8	thal	0.0033
9	thalch	0.0011
10	restecg	-0.0002

The high ranking of cp and ca illustrates the interpretability advantage of Logistic Regression, allowing direct mapping between clinical features and predictive outcomes. At the same time, the prominence of dataset origin as a top feature suggests potential cohort-specific artifacts or biases embedded within the Heart Desease dataset. Such findings underscore a dual perspective: on one hand, feature importance confirms known medical knowledge; on the other, it warns researchers of hidden dataset limitations that must be addressed before clinical deployment. This makes feature analysis a valuable tool for both validating existing theories and identifying structural issues in medical datasets.

E. Discussion and Related Work

The results of this study reinforce the importance of evaluating not only discrimination metrics but also calibration when developing predictive models for clinical applications. While many prior studies on cardiovascular risk prediction have emphasized AUROC as the principal measure of performance, our findings indicate that high AUROC does not necessarily guarantee reliable probability estimates. Logistic Regression consistently achieved balanced results across discrimination and calibration metrics, confirming its robustness for small structured medical datasets. In contrast, ensemble methods such as Random Forest and Gradient Boosting, despite their capacity to capture non-linear patterns, did not consistently deliver superior calibration. These observations are in line with several statistical reports that highlight the trade-off between complexity and interpretability, particularly in medical datasets where reliability is essential.

Furthermore, the emphasis on probability calibration directly connects this work with broader discussions in applied statistics and public health. The stability of Logistic Regression underlines its suitability for integration into clinical decision support systems, where probability thresholds guide patient management strategies. This aligns with Sustainable Development Goal (SDG) 3, which emphasizes strengthening early detection and effective management of non-communicable diseases such as cardiovascular illness. By ensuring that models provide well-calibrated probabilities, this study contributes both methodologically, by advancing applied statistical modeling practices, and practically, by supporting health systems in allocating limited resources more effectively.

IV. CONCLUSION

This study demonstrated that Logistic Regression (LR) outperformed Random Forest (RF) and Gradient Boosting (GB) when applied to the UCI Heart Disease dataset, particularly in terms of balanced discrimination and probability calibration. LR without additional calibration achieved the best overall performance with AUROC 0.903 and AUPRC 0.911, while maintaining strong calibration properties. This finding highlights that simple and interpretable models can remain highly competitive, even compared to more complex ensemble approaches, when applied to small and structured medical datasets.

The novelty of this research lies in its emphasis on probability calibration rather than solely focusing on discrimination metrics. The results show that models with slightly lower AUROC but superior calibration may be more trustworthy for clinical decision-making. Moreover, feature importance analysis confirmed the medical relevance of predictors such as chest pain type (cp), number of major vessels (ca), and ST depression (oldpeak), while also highlighting potential dataset specific artifacts.

From a practical perspective, this work contributes to Sustainable Development Goal (SDG) 3: Good Health and Well-being, by providing a calibrated modeling strategy that supports early detection of cardiovascular risk. Nevertheless, the study is limited by the relatively small and homogeneous dataset. Future research should validate these findings across larger and more diverse cohorts, and explore alternative calibration methods or hybrid modeling approaches to further enhance clinical applicability.

REFERENCES

[1] W. Adisasmito, V. Amir, A. Atin, A. Megraini, and D. Kusuma, "Geographic and socioeconomic disparity in cardiovascular risk factors in Indonesia: analysis of the Basic Health Research 2018," BMC Public Health, vol. 20, no. 1, p. 1004, Jun. 2020, doi: 10.1186/s12889-020-09099-1.

[2] S. Sujarwoto et al., "Healthcare access and socio-demographic determinants of estimated 10-year risk of cardiovascular diseases in Indonesia: A population-based study," PLOS ONE, vol. 20, no. 8, p. e0318112, Aug. 2025, doi: 10.1371/journal.pone.0318112.

[3] "Cardiovascular diseases (CVDs)." Accessed: Sep. 01, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[4] D. S. Arsyad et al., "Modifiable risk factors in adults with and without prior cardiovascular disease: findings from the Indonesian National Basic Health Research," BMC Public Health, vol. 22, no. 1, p. 660, Apr. 2022, doi: 10.1186/s12889-022-13104-0.

[5] J. Singh Thakur, R. Nangia, and S. Singh, "Progress and challenges in achieving noncommunicable diseases targets for the sustainable development goals," FASEB BioAdvances, vol. 3, no. 8, pp. 563–568, 2021, doi: 10.1096/fba.2020-00117.

[6] R. Nugent et al., "Investing in non-communicable disease prevention and management to advance the Sustainable Development Goals," The Lancet, vol. 391, no. 10134, pp. 2029–2035, May 2018, doi: 10.1016/S0140-6736(18)30667-6.

[7] S. P. Karunathilake and G. U. Ganegoda, "Secondary Prevention of Cardiovascular Diseases and Application of Technology for Early Diagnosis," BioMed Res. Int., vol. 2018, no. 1, p. 5767864, 2018, doi: 10.1155/2018/5767864.

[8] A. Kumar, Er. R. Khan, and Deepika, "A Review On Heart Disease Detection Using Machine Learning Techniques," in 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT), Apr. 2024, pp. 317–323. doi: 10.1109/CCICT62777.2024.00059.

[9] "Full article: Systematic reviews of machine learning in healthcare: a literature review." Accessed: Sep. 01, 2025. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14737167.2023.2279107

[10] L. Xu, L. Sanders, K. Li, and J. C. L. Chow, "Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review," JMIR Cancer, vol. 7, no. 4, p. e27850, Nov. 2021, doi: 10.2196/27850.

[11] K. Morooka, M. Nakamoto, and Y. Sato, "A Survey on Statistical Modeling and Machine Learning Approaches to Computer Assisted Medical Intervention: Intraoperative Anatomy Modeling and Optimization of Interventional Procedures," IEICE Trans. Inf., vol. E96-D, no. 4, pp. 784–797, Apr. 2013, doi: 10.1587/transinf.E96.D.784.

[12] E. Miranda, F. M. Bhatti, M. Aryuni, and C. Bernando, "Intelligent Computational Model for Early Heart Disease Prediction using Logistic Regression and Stochastic Gradient Descent (A Preliminary Study)," in 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), Oct. 2021, pp. 11–16. doi: 10.1109/ICCSAI53272.2021.9609724.

[13] Z. Selvitopi and H. Selvitopi, "Machine learning methods for predicting cardiovascular diseases analyzing a hybrid dataset,"

Procedia Comput. Sci., vol. 258, pp. 3535–3543, 2025, doi: 10.1016/j.procs.2025.04.609.

[14] "Model-Based ROC Curve: Examining the Effect of Case Mix and Model Calibration on the ROC Plot - Mohsen Sadatsafavi, Paramita Saha-Chaudhuri, John Petkau, 2022." Accessed: Sep. 01, 2025. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0272989X211050909

[15] A. M. Carrington et al., "Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 329–341, Jan. 2023, doi: 10.1109/TPAMI.2022.3145392.

[16] P. P. Win, S. W. Phyo, and K. K. Zaw, "Comparative Analysis of Predicting Hospitalization Time for Diabetes Patients Using Gradient Boosting and Random Forest Algorithms," in 2024 5th International Conference on Advanced Information Technologies (ICAIT), Nov. 2024, pp. 1–6. doi: 10.1109/ICAIT65209.2024.10754940.

[17] "Predicting Adult Hospital Admission from Emergency Department Using Machine Learning: An Inclusive Gradient Boosting Model." Accessed: Sep. 01, 2025. [Online]. Available: https://www.mdpi.com/2077-0383/11/23/6888

[18] B. T. Mashi, M. Hamada, J. J. Tanimu, P. Robert, and T. J. Samson, "An Ensemble Approach for Stroke Prediction," in 2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Dec. 2024, pp. 381–388. doi: 10.1109/MCSoC64144.2024.00069.

[19] P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, "From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies," Mob. Inf. Syst., vol. 2022, no. 1, p. 8167821, 2022, doi: 10.1155/2022/8167821.

[20] T. Wang and Q. Lin, "Hybrid Predictive Models: When an Interpretable Model Collaborates with a Black-box Model," J. Mach. Learn. Res., vol. 22, no. 137, pp. 1–38, 2021.

[21] A. Maalej, U. Johansson, and T. Lofstrom, "Evaluating Calibration Techniques for Reliable Predictions," in Machine Learning and Soft Computing, L. Huang, Ed., Singapore: Springer Nature, 2025, pp. 159–175. doi: 10.1007/978-981-96-6403-0_14.

[22] S. Xu, Z. Jiang, Z. Chen, D. Pan, H. Yu, and L. Li, "Blast Furnace Condition Recognizing in the Ironmaking Process Based on Prior Knowledge and Platt Scaling Probability," in 2024 IEEE International Conference on Industrial Technology (ICIT), Mar. 2024, pp. 1–6. doi: 10.1109/ICIT58233.2024.10540890.

[23] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "'Heart Desease' UCI Machine Learning Repository." doi: https://doi.org/10.24432/C52P4X.

[24] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

[25] K. K. Napa, R. Govindarajan, S. Sathya, J. S. Murugan, and B. K. P. Vijayammal, "Comparative analysis of explainable machine learning models for cardiovascular risk stratification using clinical data and shapley additive explanations," Intell.-Based Med., vol. 12, p. 100286, Jan. 2025, doi: 10.1016/j.ibmed.2025.100286.

[26] S. Tribuvan et al., "Performance Evaluation of Advanced Classification Models Combined with Feature Selection for Credit Risk Performance," Procedia Comput. Sci., vol. 258, pp. 278–287, Jan. 2025, doi: 10.1016/j.procs.2025.04.265.

[27] J. Meng and R. Xing, "Inside the 'black box': Embedding clinical knowledge in data-driven machine learning for heart disease diagnosis," Cardiovasc. Digit. Health J., vol. 3, no. 6, pp. 276–288, Dec. 2022, doi: 10.1016/j.cvdhj.2022.10.005.

[28] "Heart Disease Prediction Model Using Feature Selection and Ensemble Deep Learning with Optimized Weight," CMES - Comput. Model. Eng. Sci., vol. 143, no. 1, pp. 875–909, Apr. 2025, doi: 10.32604/cmes.2025.061623.

[29] "What Does Your Bio Say? Inferring Twitter Users' Depression Status From Multimodal Profile Information Using Deep Learning | IEEE Journals & Magazine | IEEE Xplore." Accessed: Sep. 01, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/9567734

[30] "Goal 3 | Department of Economic and Social Affairs." Accessed: Sep. 01, 2025. [Online]. Available: https://sdgs.un.org/goals/goal3?utm_source=chatgpt.com

[31] "SDG Target 3.4 | Noncommunicable diseases and mental health: By 2030, reduce by one third premature mortality from non-communicable diseases through prevention and treatment and promote mental health and well-being." Accessed: Sep. 01, 2025. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/sdg-target-3.4-noncommunicable-diseases-and-mental-health?utm_source=chatgpt.com

[32] Bayuaji L, Amzah MY, Pebrianti D. Optimization of feature selection in support vector machines (SVM) using recursive feature elimination (RFE) and particle swarm optimization (PSO) for heart disease detection. In2024 9th International Conference on Mechatronics Engineering (ICOM) 2024 Aug 13 (pp. 304-309). IEEE