p-ISSN: 2963-7856| e-ISSN: 2961-9890
Available online at jerkin.org/index.php/jerkin
Jurnal Pengabdian Masyarakat dan Riset Pendidikan
Volume 4 No. 3 Januari-Maret 2026, pp 18718-18728

Predicting Heart Failure Status Using Binary Logistic Regression with Clinical and Demographic Factors

Nita Cahyani[1*], Rahmat Irsyada[2]

[1]Universitas Padjadjaran, Jl. Raya Bandung-Sumedang Km. 21 Jatinangor, Kab. Sumedang. Jawa Barat.
[2]Politeknik Negeri Subang, Jl. Brigjen Katamso No. 37, Kec. Subang, Kab. Subang, Jawa Barat.
E-mail: nita.cahyani@unpad.ac.id
* Corresponding Author

https://doi.org/10.31004/jerkin.v4i3.5189

ARTICLE INFO

Article history
Received: 05 Dec 2025
Revised: 17 Dec 2025
Accepted: 30 Dec 2025

Kata Kunci:
Gagal Jantung, Regresi Logistik, Rasio Odds, ROC–AUC, Prediktor Klinis

Keywords:
Heart Failure, Logistic Regression, Odds Ratio, ROC–AUC, Clinical Predictors

ABSTRACT

Tujuan penelitian ini adalah mengidentifikasi karakteristik klinis dan demografis yang berhubungan dengan gagal jantung serta mengembangkan model risiko yang dapat diinterpretasikan menggunakan regresi logistik biner pada data pasien rumah sakit. Deteksi dini gagal jantung diharapkan mendukung intervensi tepat waktu dan pengambilan keputusan klinis berbasis pengukuran rutin. Penelitian ini menganalisis 130 data pasien anonim dengan status gagal jantung sebagai keluaran biner. Model regresi logistik awal mencakup seluruh prediktor kandidat, kemudian disederhanakan untuk meningkatkan stabilitas dan kalibrasi. Hasil disajikan sebagai rasio odds dengan CI 95%, evaluasi kinerja meliputi ROC–AUC, metrik klasifikasi, uji Hosmer–Lemeshow, plot kalibrasi, serta validasi silang 5-fold. Model akhir signifikan (LR p = $1{,}0\times10^{-5}$; McFadden R^2 = 0,222) dengan akurasi 81,54%, sensitivitas 89,41%, spesifisitas 66,67%, AUC 0,811, dan skor Brier 0,164. Validasi silang menunjukkan AUC rata-rata 0,774 dan akurasi 0,762. Prediktor signifikan meliputi BMI, kreatinin serum, kalium serum, dan kolesterol total, dengan kalibrasi yang dapat diterima (p = 0,0767). Model ini berpotensi digunakan sebagai alat skrining yang interpretatif, meskipun validasi eksternal masih diperlukan.

The aim of this study was to identify clinical and demographic characteristics associated with heart failure and develop an interpretable risk model using binary logistic regression on hospital patient data. Early detection of heart failure is expected to support timely intervention and clinical decision-making based on routine measurements. This study analyzed 130 anonymized patient data with heart failure status as a binary outcome. The initial logistic regression model included all candidate predictors and was then simplified to improve stability and calibration. Results are presented as odds ratios with 95% CIs. Performance evaluation included ROC–AUC, classification metrics, the Hosmer–Lemeshow test, calibration plots, and 5-fold cross-validation. The final model was significant (LR p = 1.0×10^{-5}; McFadden R^2 = 0.222) with an accuracy of 81.54%, sensitivity of 89.41%, specificity of 66.67%, AUC of 0.811, and a Brier score of 0.164. Cross-validation showed an average AUC of 0.774 and an accuracy of 0.762. Significant predictors included BMI, serum creatinine, serum potassium, and total cholesterol, with acceptable calibration (p = 0.0767). This model has potential use as an interpretive screening tool, although external validation is still needed.

How to Cite: Nita Cahyani, et al (2025). Predicting Heart Failure Status Using Binary Logistic Regression with Clinical and Demographic Factors, 4(3) 18718-18728. https://doi.org/10.31004/jerkin.v4i3.5189

INTRODUCTION

Heart failure (HF) is still a major challenge in clinical practice. Many patients experience repeated hospital admissions, their daily functioning declines, and mortality remains high, together placing ongoing strain on health systems (Heidenreich et al., 2022; Lala et al., 2025). Contemporary international guidelines increasingly emphasize a prevention-oriented framework, highlighting that a

large proportion of HF risk can be addressed earlier through recognition and management of upstream cardiovascular and metabolic factors (Heidenreich et al., 2022; Lala et al., 2025; Wei et al., 2025). This prevention focus is crucial because HF typically evolves over time, often preceded by long-standing exposure to modifiable risks that may be detectable through routinely collected clinical information (Heidenreich et al., 2022; Lala et al., 2025).

Among modifiable risk factors, elevated blood pressure has been repeatedly linked to the development and worsening of HF (Heidenreich et al., 2022). Large evidence syntheses show that blood-pressure lowering reduces major cardiovascular outcomes, including HF events, which is why tight blood-pressure control is widely viewed as a central prevention strategy (Collaboration, 2021; Ettehad et al., 2015). Excess body weight is another well-established contributor; meta-analyses consistently report higher HF risk among people who are overweight or obese, supporting the use of body mass index (BMI) in risk stratification (Aune et al., 2016; Lachute, Seltz, Lavie, & Mandras, 2024). Smoking further increases cardiovascular risk, and both cohort studies and meta-analyses indicate higher HF incidence among smokers, while smoking cessation is associated with a meaningful reduction in cardiovascular harm over time (Ding et al., 2022; Lee, 2019; Yoo et al., 2023).

HF risk assessment also draws on systemic markers that capture how multiple organs are involved. Kidney dysfunction is closely connected to HF through overlapping mechanisms and a two-way pattern of disease progression, and evidence from both community and clinical cohorts points to kidney function as a meaningful indicator of incident HF risk (Buckley et al., 2025; Zelnick et al., 2022). In addition, routine laboratory tests such as electrolytes and hemoglobin are often used in clinical profiling because they can signal disease severity, comorbidity burden, or treatment effects, and they can help produce more informative risk estimates in real-world practice (Ferreira JP, Butler J, Rossignol P, 2020; Heidenreich et al., 2022; Polcwiartek et al., 2018).

To turn routinely collected clinical information into practical decision support, clinical prediction modeling offers a structured way to estimate individual risk and guide early screening strategies (Cahyani, Fithriasari, & Iriawan, 2018; Cahyani, Pangastuti, Fithriasari, Irhamah, & Iriawan, 2021; Cahyani & Irsyada, 2025; Iriawan et al., 2018; Zhang, Golbus, Wittrup, Aaronson, & Najarian, 2024). Logistic regression is still widely used in biomedical research because it handles binary outcomes well and provides results that are straightforward to interpret through odds ratios and confidence intervals, which is especially valuable when interpretability matters (Arif & Cahyani, 2022; Collins, Reitsma, & Altman, 2015). Importantly, model assessment should not stop at identifying "significant" predictors; it also needs to report predictive performance, since discrimination and calibration describe different but equally important aspects of clinical usefulness (Alba et al., 2017).

Transparent reporting and careful risk-of-bias assessment are particularly important when the sample size is modest and the risk of overfitting cannot be ignored. The TRIPOD statement provides a clear standard for reporting how prediction models are developed and validated, and TRIPOD+AI extends this guidance to ensure consistent reporting for both regression-based and machine-learning approaches (Collins et al., 2024, 2015). In addition, PROBAST offers a structured way to judge the risk of bias and applicability of prediction model studies, supporting more credible modeling decisions and performance claims (Wolff et al., 2019).

Accordingly, this study develops an interpretable binary logistic regression model to estimate the probability of documented heart failure (HF) status using a hospital dataset of 130 patient records. HF status was extracted from the medical record as recorded by clinicians and coded as a binary outcome (1 = HF, 0 = non-HF). Predictors comprised routinely available demographic and clinical measures, including age, sex, smoking status, body mass index (BMI), systolic and diastolic blood pressure, and laboratory biomarkers such as creatinine and electrolytes. Model performance was evaluated using standard classification metrics and the ROC–AUC to assess discrimination, along with calibration-focused assessment to examine agreement between predicted probabilities and observed outcomes for potential screening-oriented use (Alba et al., 2017).

METHOD

Study design and reporting standard

This study used a quantitative, retrospective observational design based on secondary clinical data to develop and evaluate a prediction model for heart failure (HF) status. The methods are reported in

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

line with established guidance for clinical prediction modeling, with clear descriptions of the candidate predictors, the modeling approach, and the performance measures, following the TRIPOD and TRIPOD+AI recommendations (Collins et al., 2024, 2015). Considerations related to risk of bias and applicability were addressed using the PROBAST framework (Wolff et al., 2019). Model performance reporting includes both discrimination and calibration, consistent with current recommendations for evaluating prediction models (Alba et al., 2017; Calster, Mclernon, Smeden, Wynants, & Steyerberg, 2019; Riley, Snell, et al., 2024).

Data source, sample size, and outcome

The dataset was obtained from the medical record system of RSUD Dr. R. Sosodoro Djatikoesoemo, Bojonegoro, Indonesia. Records collected between August and December 2023 were extracted and de-identified prior to analysis. The final dataset included 130 patient records, with heart failure (HF) status as a binary outcome (1 = HF, 0 = non-HF). All eligible records available during the study period were included (total sampling). Candidate predictors comprised demographic characteristics and routinely recorded clinical and laboratory variables, including age, sex, employment status, education level, smoking status, marital status, body mass index (BMI), hemoglobin, total cholesterol, serum creatinine, serum sodium, serum potassium, systolic blood pressure, and diastolic blood pressure.

Variable coding and preprocessing

Binary predictors were coded as 0/1. Education was treated as a categorical variable and entered using treatment (dummy) coding, with "no schooling" as the reference category (Collins et al., 2024, 2015). Before model fitting, variables recorded as strings were converted to numeric values where appropriate. Basic data checks were conducted, including screening for missing values and verifying that numeric entries fell within plausible ranges. No missing values were found in the analysis variables; therefore, all analyses were performed using complete-case data (n = 130).

Model development: binary logistic regression

Binary logistic regression was used to model heart failure (HF) status as a dichotomous outcome. As an initial exploratory step, we fitted a full model that included all candidate predictors available in the dataset (demographic characteristics and routine clinical and laboratory measurements). This full specification was used as a sensitivity analysis to describe the direction and strength of associations in the complete set of variables.

Given the modest sample size relative to the number of candidate predictors, and recognizing that prediction models should be judged not only by discrimination but also by calibration, we then defined a reduced model as the primary prediction model. The reduced model retained clinically plausible and routinely available predictors and included age, sex, smoking status, body mass index (BMI), total cholesterol, serum creatinine, serum potassium, and diastolic blood pressure. This parsimonious approach was intended to improve numerical stability, reduce the risk of overfitting, and provide more reliable probability estimates for screening-oriented use.

All models were estimated using maximum likelihood. Model results are reported as odds ratios (ORs) with 95% confidence intervals, alongside p-values from Wald tests. Logistic regression was selected because it is widely used for binary clinical outcomes and yields effect estimates that are straightforward to interpret, which is advantageous when model transparency is important (Alba et al., 2017; Collins et al., 2024).

Overall model fit, discrimination, and calibration

Model evaluation followed current recommendations that clinical prediction models should be assessed across three complementary domains: overall fit, discrimination, and calibration (Alba et al., 2017; Calster et al., 2019; Collins et al., 2024; Riley, Archer, et al., 2024).

1. Overall fit: Global model significance was evaluated using the likelihood ratio test comparing the fitted model with the intercept-only model. Model fit and parsimony were further summarized using the deviance (−2 log-likelihood) and information criteria (AIC and BIC).
2. Discrimination: Discrimination was assessed using receiver operating characteristic (ROC) analysis and the area under the curve (AUC) (Alba et al., 2017; Carter, Pan, Rai, & Galandiuk, 2016). In addition, threshold-based classification performance at a probability cut-off of 0.50 was summarized using the confusion matrix and derived measures including sensitivity (true positive rate) and

specificity (true negative rate). The AUC was used as the primary summary of discrimination because it reflects the model's ability to rank patients by risk across all possible thresholds.

3. Calibration: Calibration, defined as agreement between predicted probabilities and observed outcomes, was examined using the Hosmer–Lemeshow goodness-of-fit test based on deciles of predicted risk. This was complemented by a 10-bin calibration plot comparing mean predicted probability with the observed event rate within each bin (Calster et al., 2019; Hosmer & Lemesbow, 1980; Surjanovic & Loughin, 2024).

Note on validation: Because model performance was first assessed on the same dataset used for model development, these results reflect apparent (in-sample) performance and may be optimistic. To quantify optimism and stability, internal validation was performed using stratified 5-fold cross-validation for the primary reduced model (Collins et al., 2024; Riley, Snell, et al., 2024).

Classification metrics and decision threshold

Predicted probabilities from the logistic regression model were converted into binary class labels using a fixed decision threshold of 0.50: patients with predicted probability ≥ 0.50 were classified as heart failure (HF = 1), and those with predicted probability < 0.50 were classified as non-HF (HF = 0). Classification performance was summarized using the confusion matrix (true positives, false positives, false negatives, and true negatives) and standard derived metrics, including accuracy, precision, sensitivity (recall), specificity, and the F1-score (Alba et al., 2017; Rainio, 2024).

As an additional sensitivity analysis, an alternative "optimal" probability threshold was determined from the ROC curve using Youden's index, which selects the cut-point that provides the best balance between sensitivity and specificity (Carter et al., 2016). The resulting threshold and its corresponding classification metrics were reported.

Multicollinearity assessment

Potential multicollinearity among predictors was evaluated using the variance inflation factor (VIF), calculated from the fitted design matrix. VIF values quantify how strongly each predictor is linearly explained by the other predictors; larger values indicate greater collinearity. As a practical guideline, VIF values above 5 may suggest moderate concern and values above 10 may indicate substantial multicollinearity (Huang, Jou, & Cho, 2016).

Software

All analyses were conducted in Python, using libraries for data handling and statistical modeling (pandas, numpy, statsmodels), evaluation metrics (scikit-learn), statistical testing (scipy), visualization (matplotlib), and Excel output generation (openpyxl) (Collins et al., 2024; Riley, Archer, et al., 2024).

RESULTS AND DISCUSSION

Participant Characteristics

A total of 130 patients were included in the analysis. Of these, 85 (65.4%) were classified as having heart failure (HF = 1) and 45 (34.6%) as non-heart failure (HF = 0). The mean age was 58.68 ± 12.73 years (range: 19–86), and the mean body mass index (BMI) was 21.60 ± 2.46 kg/m². Mean systolic and diastolic blood pressures were 129.22 ± 24.52 mmHg and 77.43 ± 15.48 mmHg, respectively. Among categorical characteristics, 72 patients (55.4%) were male, 62 (47.7%) were smokers, 107 (82.3%) were employed, and 82 (63.1%) were married. Baseline characteristics are summarized in Table 1.

Table 1. Baseline characteristics of study participants (n = 130)

Characteristic	Overall (n = 130)
Heart failure status	
HF (Y=1)	85 (65.4%)
Non-HF (Y=0)	45 (34.6%)
Age (years), mean ± SD	58.68 ± 12.73
Range (min–max)	19–86
Body mass index (kg/m²), mean ± SD	21.60 ± 2.46
Systolic blood pressure (mmHg), mean ± SD	129.22 ± 24.52
Diastolic blood pressure (mmHg), mean ± SD	77.43 ± 15.48

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Male, n (%)	72 (55.4%)
Smoker, n (%)	62 (47.7%)
Employed, n (%)	107 (82.3%)
Married, n (%)	82 (63.1%)

Primary prediction model: reduced logistic regression

Given the modest sample size relative to the number of candidate predictors and the importance of probability calibration for screening-oriented use, we selected a reduced logistic regression model as the primary prediction model. This model included age, sex, smoking status, BMI, total cholesterol, creatinine, potassium, and diastolic blood pressure.

The model converged without difficulty (LL = −65.232; LL-null = −83.854) and was statistically significant overall (LR $p = 1.0\times10^{-5}$), with McFadden's pseudo-$R^2 = 0.222$. Calibration was acceptable (Hosmer–Lemeshow $\chi^2 = 14.202$, df = 8, p = 0.0767). In terms of discrimination, the model achieved an AUC of 0.8107, indicating good ability to distinguish HF from non-HF cases. Prediction error was modest (Brier score = 0.1635). At a 0.50 threshold, the reduced model produced TN = 30, FP = 15, FN = 9, and TP = 76, corresponding to an accuracy of 0.8154, sensitivity of 0.8941, and specificity of 0.6667. Precision was 0.8352 with an F1-score of 0.8636. Internal validation using 5-fold stratified cross-validation showed stable performance, with mean AUC = 0.7739 (SD = 0.0225) and mean accuracy = 0.7615 at a 0.50 threshold.

Table 2. Summary of primary reduced model performance

Performance measure	Value
ROC–AUC (apparent, in-sample)	0.8107
Brier score (apparent, in-sample)	0.1635
Hosmer–Lemeshow p-value	0.0767
Accuracy (threshold = 0.50)	0.8154
Sensitivity / Recall (threshold = 0.50)	0.8941
Specificity (threshold = 0.50)	0.6667
5-fold CV mean AUC (SD)	0.7739 (0.0225)

Effects of Predictors in the Reduced Model (Odds Ratios)

To support interpretability, coefficients and odds ratios (ORs) with 95% confidence intervals are presented in Table 3. BMI, creatinine, potassium, and total cholesterol were statistically significant predictors in the reduced model.

Table 3. Reduced model coefficients and odds ratios (primary prediction model)

Predictor	β	OR	95% CI (OR)	p-value
Age (years)	0.0135	1.014	0.978–1.051	0.461
Sex (male=1)	0.179	1.196	0.176–8.119	0.854
Smoking (smoker=1)	−1.704	0.182	0.027–1.228	0.080
BMI (kg/m²)	0.345	1.413	1.152–1.731	0.00088
Total cholesterol	−0.00612	0.994	0.989–0.999	0.0233
Creatinine	1.347	3.845	1.427–10.357	0.00773
Potassium	0.00737	1.007	1.001–1.013	0.0141
Diastolic BP (mmHg)	−0.0275	0.973	0.946–1.001	0.0577

Confusion Matrix and ROC Curve (Reduced Model)

The confusion matrix for the reduced model at a 0.50 threshold is shown in Table 4. Figure 1 displays the ROC curve, confirming good discrimination (AUC = 0.8107). Discrimination reflects how well the model ranks patients by risk; therefore, it should be interpreted together with calibration results.

Table 4. Confusion matrix (threshold = 0.50; reduced model)

Actual / Predicted	Non-HF (0)	HF (1)	Total
Non-HF (0)	TN = 30	FP = 15	45
HF (1)	FN = 9	TP = 76	85
Total	39	91	130

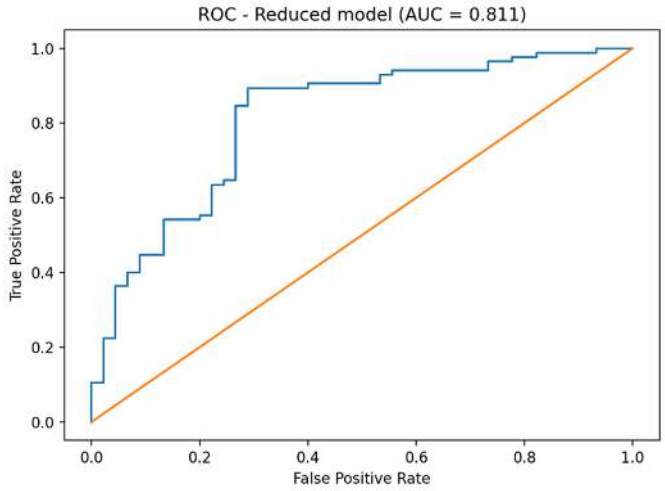

Figure 1. ROC curve of the reduced model (primary analysis)

Bottom of Form Sensitivity Analysis (Full Model)

Full Logistic Regression Model: Overall Fit, Performance, and Diagnostics

As a sensitivity (exploratory) analysis, we also fitted a full logistic regression model including all candidate predictors (age, sex, employment status, education, smoking status, marital status, BMI, hemoglobin, total cholesterol, creatinine, sodium, potassium, systolic blood pressure, and diastolic blood pressure). The model converged after 8 MLE iterations (LL = −58.477; LL-null = −83.854) and was statistically significant overall (LR p = 3.222×10^{-5}), with McFadden's pseudo-R2 = 0.3026 (Table 5).

The full model showed good apparent discrimination (AUC ≈ 0.849; Figure 2). Using a 0.50 threshold, it correctly classified 33 non-HF cases (TN) and 78 HF cases (TP), with 12 false positives and 7 false negatives (Table 7). This corresponded to an accuracy of 0.8538, sensitivity of 0.9176, specificity of 0.7333, precision of 0.8667, and an F1-score of 0.8914 (Table 8). Coefficient estimates and odds ratios are reported in Table 6, and several predictors, including smoking status, BMI, total cholesterol, creatinine, potassium, and diastolic blood pressure, showed statistically significant associations in this fitted model.

Despite these results, calibration was poor (Hosmer–Lemeshow p = 0.0020), indicating that predicted probabilities did not align well with observed event rates across risk groups. For this reason, the full model is presented as a sensitivity analysis, while the reduced model is retained as the primary prediction model due to its more acceptable calibration and better suitability for screening-oriented probability estimation.

Table 5. Overall model fit (Binary Logistic Regression)

Fit statistic	Value
Number of observations (n)	130
Degrees of freedom (model), df_model	17
Degrees of freedom (residual), df_resid	112
Log-likelihood (LL)	−58.477
Null log-likelihood (LL-null)	−83.854
Likelihood ratio test (LLR) p-value	3.222×10^{-5}
−2 Log-likelihood (−2LL)	116.953
Akaike Information Criterion (AIC)	152.953
Bayesian Information Criterion (BIC)	204.569
McFadden's pseudo-R2	0.303
Hosmer–Lemeshow χ^2(df=8)	24.336
Hosmer–Lemeshow p-value	0.0020

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Table 6. Logistic regression coefficients and decisions (α = 0.05)
(Education reference = 0 "no schooling")

Predictor	β	OR	95% CI (OR)	p-value	Decision
Education: SD (1)	−0.245	0.783	0.199–3.082	0.726	Not significant
Education: SMP (2)	1.811	6.114	0.765–48.844	0.088	Not significant (borderline)
Education: SMA (3)	0.022	1.022	0.161–6.478	0.981	Not significant
Education: D3/D4/S1 (4)	0.136	1.146	0.106–12.364	0.911	Not significant
Age (years)	0.032	1.032	0.984–1.084	0.197	Not significant
Sex (male=1)	−0.127	0.880	0.123–6.322	0.899	Not significant
Employment (working=1)	1.149	3.155	0.832–11.963	0.091	Not significant (borderline)
Smoking (smoker=1)	**−1.938**	**0.144**	**0.021–0.990**	**0.049**	**Significant**
Marital status (married=1)	−0.414	0.661	0.225–1.941	0.451	Not significant
BMI (kg/m²)	**0.398**	**1.488**	**1.161–1.907**	**0.002**	**Significant**
Hemoglobin (g/dL)	−0.044	0.957	0.778–1.178	0.679	Not significant
Total cholesterol	**−0.008**	**0.992**	**0.986–0.998**	**0.014**	**Significant**
Creatinine	**1.528**	**4.610**	**1.593–13.339**	**0.005**	**Significant**
Sodium	−0.116	0.891	0.777–1.021	0.096	Not significant (borderline)
Potassium	**0.007**	**1.007**	**1.000–1.014**	**0.043**	**Significant**
Systolic BP	0.023	1.023	0.994–1.053	0.125	Not significant
Diastolic BP	**−0.054**	**0.947**	**0.904–0.993**	**0.024**	**Significant**

Table 7. Confusion Matrix (Threshold = 0.50): full model (sensitivity analysis)

Actual / Predicted	Non-HF (0)	HF (1)	Total
Non-HF (0)	**TN = 33**	**FP = 12**	45
HF (1)	**FN = 7**	**TP = 78**	85
Total	40	90	**130**

Table 8. Classification Performance Metrics: full model (sensitivity analysis)

Metric	Value
Accuracy	0.8538
Precision	0.8667
Recall / Sensitivity	0.9176
Specificity	0.7333
F1-score	0.8914

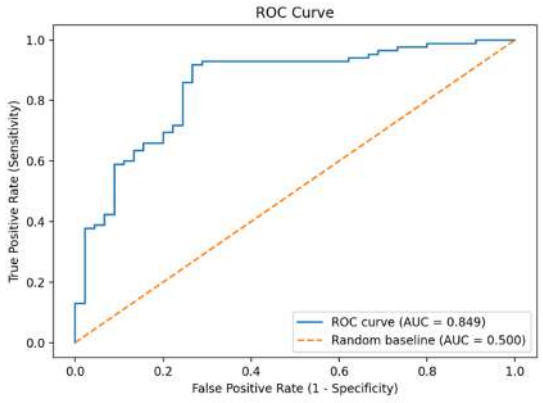

Figure 2. ROC curve of the full model (sensitivity analysis)

Calibration and Hosmer–Lemeshow test

Calibration was assessed using the Hosmer–Lemeshow (HL) goodness-of-fit test with predicted-risk deciles. The HL statistic was $\chi^2 = 24.336$ (df = 8, p = 0.0020), which indicates a statistically significant lack of fit. In practical terms, this suggests that the predicted probabilities do not align closely with the observed heart failure event rates across the risk groups, providing evidence of miscalibration.

Multicollinearity (VIF)

Multicollinearity was examined using the variance inflation factor (VIF). The largest VIF values among the predictors were around 4.4 (excluding the intercept), and systolic and diastolic blood pressure showed VIF values of approximately 2.4 to 2.5. Using common reference thresholds (VIF > 5 as moderate concern and VIF > 10 as strong concern), these results do not indicate serious multicollinearity, suggesting that the coefficient estimates are reasonably stable.

Discussion

This study developed and evaluated an interpretable logistic regression model to predict heart failure (HF) status using routinely available demographic, clinical, and laboratory variables from a hospital cohort. Because probability calibration and model stability are important when predictions may support screening, the reduced model is emphasized as the primary prediction model, while the full model is reported as a sensitivity analysis.

Key predictors in the primary reduced model

BMI was positively associated with HF status (β = 0.345; OR = 1.413; p = 0.00088). After adjustment for other predictors, each 1-unit increase in BMI was associated with approximately 41% higher odds of HF classification. This finding is consistent with evidence linking higher body weight to increased HF risk and supports BMI as a practical component for screening-oriented risk stratification.

Creatinine showed the strongest positive association (β = 1.347; OR = 3.845; p = 0.00773), indicating substantially higher odds of HF among patients with higher creatinine levels. This result is clinically plausible given the close relationship between cardiac function and kidney function. In hospital cohorts, creatinine may also reflect overall illness burden, which can contribute to its predictive value.

Total cholesterol showed a small but statistically significant inverse association with HF status (β = −0.00612; OR = 0.994; p = 0.0233). The effect per recorded unit is modest, and this direction should not be interpreted as a protective causal effect. In HF cohorts, inverse associations have been discussed as cohort-specific patterns related to severity, inflammation, nutritional status, or treatment.

Potassium was positively associated with HF status (β = 0.00737; OR = 1.007; p = 0.0141). Although the per-unit effect is small, potassium remains clinically relevant because electrolyte abnormalities are common in HF, particularly in the setting of renal dysfunction and HF therapies. Interpretation should consider the laboratory unit and scaling used in the hospital system. Future studies should harmonize laboratory units and confirm measurement conventions to strengthen clinical interpretability.

Borderline predictors and clinical interpretation

Smoking status showed borderline evidence (β = −1.704; OR = 0.182; p = 0.080) with an inverse direction. Because this direction differs from typical epidemiologic evidence, it may reflect confounding (for example, smokers being younger), selection effects, treatment differences, or coding and measurement factors. Simple exploratory checks, such as cross-tabulations of smoking status by HF status and age, are recommended before drawing clinical conclusions.

Diastolic blood pressure also showed borderline evidence (β = −0.0275; OR = 0.973; p = 0.0577). In clinical populations, lower diastolic pressure can reflect arterial stiffness and wider pulse pressure, which may be associated with higher cardiovascular risk in older or comorbid patients. However, given the borderline significance and the complexity of blood pressure dynamics and treatment patterns, this association should be interpreted primarily in predictive rather than causal terms.

Model performance, calibration, and intended use

Overall, the reduced model demonstrated good discrimination (AUC = 0.8107). Using a 0.50 threshold, it achieved accuracy = 0.8154, sensitivity = 0.8941, specificity = 0.6667, and F1-score =

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

0.8636. The high sensitivity suggests the model is useful for identifying potential HF cases, while the moderate specificity indicates that some false positives should be expected and would require follow-up assessment.

Importantly, the reduced model showed acceptable calibration (Hosmer–Lemeshow p = 0.0767), indicating better agreement between predicted probabilities and observed outcomes than the full model. Internal validation using 5-fold stratified cross-validation showed stable discrimination (mean AUC = 0.7739; SD = 0.0225), suggesting reduced optimism and supporting use of the model for screening-oriented risk ranking rather than definitive diagnosis.

Sensitivity analysis: full model

The full model achieved higher apparent discrimination (AUC ≈ 0.849) but showed poor calibration (Hosmer–Lemeshow p = 0.002), indicating that predicted probabilities were not reliable across risk strata. This pattern is consistent with overfitting risk when many predictors are included in a modest sample. Therefore, the reduced model is preferred for primary reporting and interpretation.

Limitations and future work

This study used a single-center retrospective dataset with a modest sample size and evaluation based primarily on internal validation. External validation in independent cohorts is needed before broader clinical use. In addition, laboratory unit harmonization, particularly for potassium, should be addressed to improve interpretability and transportability. Future work may also consider incorporating additional clinical information such as comorbidities, medication use, and HF subtype to further refine performance and calibration.

CONCLUSION

This study developed an interpretable reduced logistic regression model to predict heart failure status using 130 hospital patient records. The reduced model demonstrated good discrimination (AUC = 0.8107), acceptable calibration (Hosmer-Lemeshow p = 0.0767), and stable internal validation performance (5-fold cross-validated mean AUC = 0.7739, SD = 0.0225). BMI and serum creatinine showed the strongest associations with heart failure status, while total cholesterol and serum potassium also contributed significantly. Smoking status and diastolic blood pressure showed borderline evidence and should be interpreted cautiously in this hospital cohort. Overall, the model is best suited for screening-oriented risk ranking rather than definitive diagnosis. Future studies should prioritize external validation in larger and more diverse cohorts, refine calibration if needed, and harmonize laboratory units, particularly for potassium, before broader clinical implementation.

ACKNOWLEDGMENTS

Peneliti menyampaikan ucapan terima kasih kepada pihak yang sudah berkontribusi dalam pelaksanaan penelitian dan penyusunan artikel ini.

REFERENCE

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P. J., … Guyatt, G. (2017). Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA, 318(14), 1377–1384. Retrieved from https://doi.org/10.1001/jama.2017.12126

Arif, M., & Cahyani, N. (2022). Pemodelan Regresi Logistik Ordinal Pada Indeks Pembangunan Manusia (IPM) Di Jawa Timur Tahun 2020, 1(2), 64–73.

Aune, D., Sen, A., Norat, T., Janszky, I., Romundstad, P., Tonstad, S., & Vatten, L. J. (2016). Body Mass Index, Abdominal Fatness, and Heart Failure Incidence and Mortality. Circulation, 133(7), 639–649. Retrieved from https://doi.org/10.1161/CIRCULATIONAHA.115.016801

Buckley, L. F., Dorbala, P., Lamberson, V., Claggett, B. L., Ren, Y., Grams, M. E., … Shah, A. M. (2025). Linking Chronic Kidney Disease to Incident Heart Failure and Adverse Cardiac Remodeling Through the Plasma Proteome: The ARIC Study. JACC: Heart Failure, 13(8), 102512. Retrieved from https://doi.org/https://doi.org/10.1016/j.jchf.2025.102512

Cahyani, N., Fithriasari, K., & Iriawan, N. (2018). On the Comparison of Deep Learning Neural Network and Binary Logistic Regression for Classifying the Acceptance Status of Bidikmisi Scholarship Applicants in East Java Methods, 83–90.

Cahyani, N., & Irsyada, R. (2025). Performance Comparison of SelectKBest and Permutation Importance in Feature Selection for Diabetes Prediction, 5(1), 529–541.

Cahyani, N., Pangastuti, S., Fithriasari, K., Irhamah, I., & Iriawan, N. (2021). Classification of Bidikmisi Scholarship Acceptance using Neural Network Based on Hybrid Method of Genetic Algorithm. Indonesian Journal of Statistics and Its Applications, 5, 396–404. Retrieved from https://doi.org/10.29244/ijsa.v5i2p396-404

Calster, B. Van, Mclernon, D. J., Smeden, M. Van, Wynants, L., & Steyerberg, E. W. (2019). Calibration : the Achilles heel of predictive analytics, 1–7.

Carter, J. V, Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery, 159(6), 1638–1645. Retrieved from https://doi.org/https://doi.org/10.1016/j.surg.2015.12.029

Collaboration, B. P. L. T. T. (2021). Pharmacological blood pressure lowering for primary and secondary prevention of cardiovascular disease across different levels of blood pressure: an individual participant-level data meta-analysis. Lancet (London, England), 397(10285), 1625–1636. Retrieved from https://doi.org/10.1016/S0140-6736(21)00590-0

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. Van, … Ordish, J. (2024). RESEARCH METHODS AND REPORTING TRIPOD + AI statement : updated guidance for reporting clinical prediction models that use regression or machine learning methods The TRIPOD (Transparent Reporting of a multivariable prediction model for. Retrieved from https://doi.org/10.1136/bmj-2023-078378

Collins, G. S., Reitsma, J. B., & Altman, D. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, 7594(January), 1–9. Retrieved from https://doi.org/10.1136/bmj.g7594

Ding, N., Shah, A. M., Blaha, M. J., Chang, P. P., Rosamond, W. D., & Matsushita, K. (2022). Cigarette Smoking, Cessation, and Risk of Heart Failure With Preserved and Reduced Ejection Fraction, 79(23). Retrieved from https://doi.org/10.1016/j.jacc.2022.03.377

Ettehad, D., Emdin, C. A., Kiran, A., Anderson, S. G., Callender, T., Emberson, J., … Rodgers, A. (2015). Blood pressure lowering for prevention of cardiovascular disease and death : a systematic review and meta-analysis. The Lancet, 6736(15), 1–11. Retrieved from https://doi.org/10.1016/S0140-6736(15)01225-8

Ferreira JP, Butler J, Rossignol P, D. (2020). Abnormalities of Potassium in Heart Failure, 75(22). Retrieved from https://doi.org/10.1016/j.jacc.2020.04.021

Heidenreich, P. A., Bozkurt, B., Aguilar, D., Allen, L. A., Colvin, M. M., Deswal, A., … Yancy, C. W. (2022). 2022 AHA / ACC / HFSA Guideline for the Management of Heart Failure. Retrieved from https://doi.org/10.1016/j.jacc.2021.12.012

Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods, 9(10), 1043–1069. Retrieved from https://doi.org/10.1080/03610928008827941

Huang, C.-C. L., Jou, Y.-J., & Cho, H.-J. (2016). A new multicollinearity diagnostic for generalized linear models. Journal of Applied Statistics, 43(11), 2029–2043. Retrieved from https://doi.org/10.1080/02664763.2015.1126239

Iriawan, N., Fithriasari, K., Ulama, B. S. S., Suryaningtyas, W., Pangastuti, S. S., Cahyani, N., & Qadrini, L. (2018). On The Comparison: Random Forest, SMOTE-Bagging, and Bernoulli Mixture to Classify Bidikmisi Dataset in East Java. In 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM) (pp. 137–141). Retrieved from https://doi.org/10.1109/CENIM.2018.8711035

Lachute, C. L., Seltz, J., Lavie, C. J., & Mandras, S. A. (2024). Obesity and Weight Loss Strategies for Patients With Heart Failure, 12(9). Retrieved from https://doi.org/10.1016/j.jchf.2024.06.006

Lala, A., Beavers, C., Blumer, V., Brewer, L., Oliveira-gomes, D. De, Dunbar, S., … Gulati, M. (2025). American Journal of Preventive Cardiology The continuum of prevention and heart failure in cardiovascular medicine : A joint scientific statement from the Heart Failure Society of America and the American Society for Preventive Cardiology. American Journal of Preventive Cardiology, 24(August), 101069. Retrieved from https://doi.org/10.1016/j.ajpc.2025.101069

Lee, H. (2019). Influence of Smoking Status on Risk of Incident Heart Failure : A Systematic Review

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

and Meta-Analysis of Prospective Cohort Studies.

Polcwiartek, C., Hansen, S. M., Kragholm, K., Krogager, M. L., Aldahl, M., Køber, L., … Søgaard, P. (2018). Prognostic role of serum sodium levels across different serum potassium levels in heart failure patients: A Danish register-based cohort study. International Journal of Cardiology, 272, 244–249. Retrieved from https://doi.org/https://doi.org/10.1016/j.ijcard.2018.08.045

Rainio, O. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 1–14. Retrieved from https://doi.org/10.1038/s41598-024-56706-x

Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., … Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): how to undertake an external validation study, (part 2), 1–12. Retrieved from https://doi.org/10.1136/bmj-2023-074820

Riley, R. D., Snell, K. I. E., Archer, L., Ensor, J., Debray, T. P. A., Calster, B. Van, … Collins, G. S. (2024). Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study, (part 3). Retrieved from https://doi.org/10.1136/bmj-2023-074821

Surjanovic, N., & Loughin, T. M. (2024). Improving the Hosmer-Lemeshow goodness-of-fit test in large models with replicated Bernoulli trials. Journal of Applied Statistics, 51(7), 1399–1411. Retrieved from https://doi.org/10.1080/02664763.2023.2272223

Wei, Y., Liu, S., Mu, Y., Liang, X., Chen, Z., Liu, Y., & Dong, G. (2025). Prediction models and risk scores in different types of heart failure : a review, (November). Retrieved from https://doi.org/10.3389/fmed.2025.1652307

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., … Mallett, S. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine, 170(1), 51–58. Retrieved from https://doi.org/10.7326/M18-1376

Yoo, J. E., Jeong, S.-M., Yeo, Y., Jung, W., Yoo, J., Han, K., … Shin, D. W. (2023). Smoking Cessation Reduces the Risk of Heart Failure: A Nationwide Cohort Study. JACC. Heart Failure, 11(3), 277–287. Retrieved from https://doi.org/10.1016/j.jchf.2022.07.006

Zelnick, L. R., Shlipak, M. G., Soliman, E. Z., Anderson, A., Christenson, R., Kansal, M., … Bansal, N. (2022). Prediction of Incident Heart Failure in CKD: The CRIC Study. Kidney International Reports, 7(4), 708–719. Retrieved from https://doi.org/10.1016/j.ekir.2022.01.1067

Zhang, Y., Golbus, J. R., Wittrup, E., Aaronson, K. D., & Najarian, K. (2024). Enhancing heart failure treatment decisions : interpretable machine learning models for advanced therapy eligibility prediction using EHR data. BMC Medical Informatics and Decision Making, 1–14. Retrieved from https://doi.org/10.1186/s12911-024-02453-y