MEDICINE SOUTHEAST ASIA

SOPHISTICA
WHERE INNOVATION IN MEDICINE BEGINS

Article

Clinically useful and reliable mortality scoring based on serum creatinine, ejection fraction, and age in Heart Failure: A logistic regression cost–benefit analysis

Muhammad Iqhrammullah[1*], Derren Rampengan[2], Jade Rampengan[3], Starry H Rampengan[4]

[1] Postgraduate Program of Public Health, Universitas Muhammadiyah Aceh, Banda Aceh, Indonesia
[2] Faculty of Medicine, Universitas Sam Ratulangi, Manado, Indonesia
[3] Faculty of Medicine, Universitas Katolik Atma Jaya, Jakarta, Indonesia
[4] Division of Interventional Cardiology, Department of Cardiology and Vascular Medicine, Faculty of Medicine, Universitas Sam Ratulangi - R.D. Kandou General Hospital, Manado, Indonesia
* Correspondence: Muhammad Iqhrammullah (email: iptnholdings@gmail.com)

Abstract

Background: Accurate prediction of mortality in heart failure (HF) is essential for timely clinical decision-making. However, many risk models rely on multiple inputs that may be impractical in routine settings. This study evaluated whether a simplified logistic regression model using serum creatinine, ejection fraction, and age could provide predictive performance and clinical utility comparable to more complex models.

Results: Model 2, which included serum creatinine, ejection fraction, and age, demonstrated strong performance (AUC = 0.786; Brier score = 0.166) with excellent calibration (Hosmer–Lemeshow p = 0.765; Emax = 0.064) and the highest net clinical benefit across threshold probabilities ranging from 0.10 to 0.60. Compared to Model 1, it significantly improved patient classification (NRI = 0.111; 95% CI: 0.009–0.210), while the addition of serum sodium (Model 3) or other predictors (Model 4) yielded no further reclassification gains (NRI = 0.000). Model 2 stratified patients into three predicted risk groups: low (≤33%), moderate (34–66%), and high (≥67%). This score-based classifier achieved 92.3% overall accuracy with excellent diagnostic performance across categories (e.g., High-risk: sensitivity = 1.00; specificity = 0.93).

Methods: We performed a secondary analysis of a publicly available dataset of 299 HF patients from two centers. Four logistic regression models were developed: Model 1 (serum creatinine and ejection fraction), Model 2 (Model 1 plus age), Model 3 (Model 2 plus serum sodium), and Model 4 (a full model including all available predictors: age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, and smoking). Model performance was assessed using AUC for discrimination, Brier score, and Hosmer–Lemeshow test for calibration, and decision curve analysis (DCA) for clinical utility. Net Reclassification Improvement (NRI) was used to quantify reclassification gains. A point-based scoring system was derived from the best model using age as the reference to weight predictors and estimate individual mortality risk. Patients were grouped into low, moderate, or high-risk criteria, where a confusion matrix was used to evaluate sensitivity, specificity, and predictive values for each category.

Academic Editor: Maulana A. Empitu

Received: 10 October 2025
Revised: 25 October 2025
Accepted: 11 December 2025
Published: date

Citation: To be added by editorial staff during production.

Conclusion: A simplified logistic regression model including only serum creatinine, ejection fraction, and age provides comparable predictive accuracy and greater clinical utility than a full model using 11 predictors. These findings support its use as a practical and low-cost tool for early risk stratification in heart failure care.

Keywords: Cardiac failure; scoring system; risk stratification; decision curve analysis; net reclassification improvement

1. Introduction

Heart failure (HF) remains a leading cause of morbidity and mortality worldwide, particularly among older adults and those with reduced left ventricular systolic function [1,2]. A study reported that from 1990 to 2021, global HF cases in older adults rose from 14.1 to 36.2 million, with the fastest growth observed in low- to middle-income countries (LMICs) [3]. This trend, driven largely by population growth and ageing, is projected to continue through 2035 [3]. Early identification of patients at high risk of adverse outcomes is critical for guiding treatment decisions, allocating healthcare resources, and improving survival. Traditional risk scores often require a wide array of clinical inputs—such as echocardiographic measurements, serum biomarkers, and detailed laboratory values—that may not be routinely available or feasible to collect in low-resource settings [4,5].

With the growing accessibility of electronic health records and computational tools, machine learning and artificial intelligence (AI) offer promising opportunities to simplify risk prediction without compromising accuracy. A review emphasized how AI can revolutionize cardiac care by streamlining workflows and enhancing clinical efficiency [6]. The utility of AI in supporting decision-making and improving care delivery in cardiology has been widely acknowledged among scholar communities [7,8]. However, in many LMICs, digital infrastructure remains limited, and even basic data inputs may not be consistently available [9]. Even within the machine learning paradigm, parsimonious models—those using the fewest necessary predictors—are often more favorable for practical deployment due to their improved interpretability, reduced data requirements, faster computation, and greater feasibility in real-world clinical workflows [10,11]. Therefore, in this study, we focus on developing a simple and interpretable risk scoring system based on a parsimonious set of predictors, dedicated for use in resource-limited settings.

Previous research by Chicco and Jurman demonstrated that serum creatinine and ejection fraction alone could yield reasonable predictive performance for survival in heart failure patients [12]. However, the clinical deployability of such minimal models has not been fully evaluated in terms of calibration, clinical net benefit, and reclassification accuracy. In this study, we compare multiple logistic regression models—including both minimal and full-feature variants—using rigorous model assessment techniques such as calibration analysis, decision curve analysis (DCA), and Net Reclassification Improvement (NRI). Our aim is to determine whether a simple, interpretable model can offer comparable performance to more complex models and serve as a practical tool for early risk stratification in heart failure management. To facilitate clinical implementation, we further transformed the best-performing model into a point-based scoring system by scaling regression coefficients relative to age, allowing clinicians to estimate individual risk using readily available patient data without the need for complex calculations or digital tools.

2. Results

2.1. Discrimination and calibration

The calibration performance of the four logistic regression models using both visual and statistical metrics is presented in **Figure 1** and **Table 1**. Model 1, which included only serum creatinine and ejection fraction, achieved an area under the ROC curve (AUC) of 0.761, a Brier score of 0.178, and a non-significant Hosmer–Lemeshow (HL) test p-value of 0.691—indicating acceptable discrimination and reasonable calibration. However, calibration plots revealed overestimation of risk at higher predicted probabilities (Emax = 0.205), suggesting some deviation from perfect alignment between predicted and observed outcomes.

Adding age in Model 2 improved both discrimination and calibration. The AUC increased to 0.786, the Brier score decreased to 0.166, and the HL test p-value rose to 0.765. Most notably, the Emax value decreased to 0.064, indicating closer alignment to the ideal prediction line and better overall calibration across risk strata. Models 3 and 4 did not meaningfully improve performance beyond Model 2. Both models retained identical AUC values (0.786) and similar Brier scores (~0.166) and HL p-values (~0.765), with no gain in calibration accuracy (Emax ≈ 0.064). These findings confirm that additional predictors contributed little added value, reinforcing the clinical efficiency of Model 2.

Table 1. Performance comparison of logistic regression models using discrimination, calibration, and overall accuracy metrics

Model Comparison	Predictors Included	AUC (C-stat)	Brier Score	HL p-value	Emax
Model 1	Serum creatinine + ejection fraction	0.761	0.178	0.691	0.205
Model 2	+ age	0.786	0.166	0.765	0.064
Model 3	+ serum_sodium	0.786	0.166	0.765	0.064
Model 4	All variables	0.786	≈ 0.166	≈ 0.765	≈ 0.064

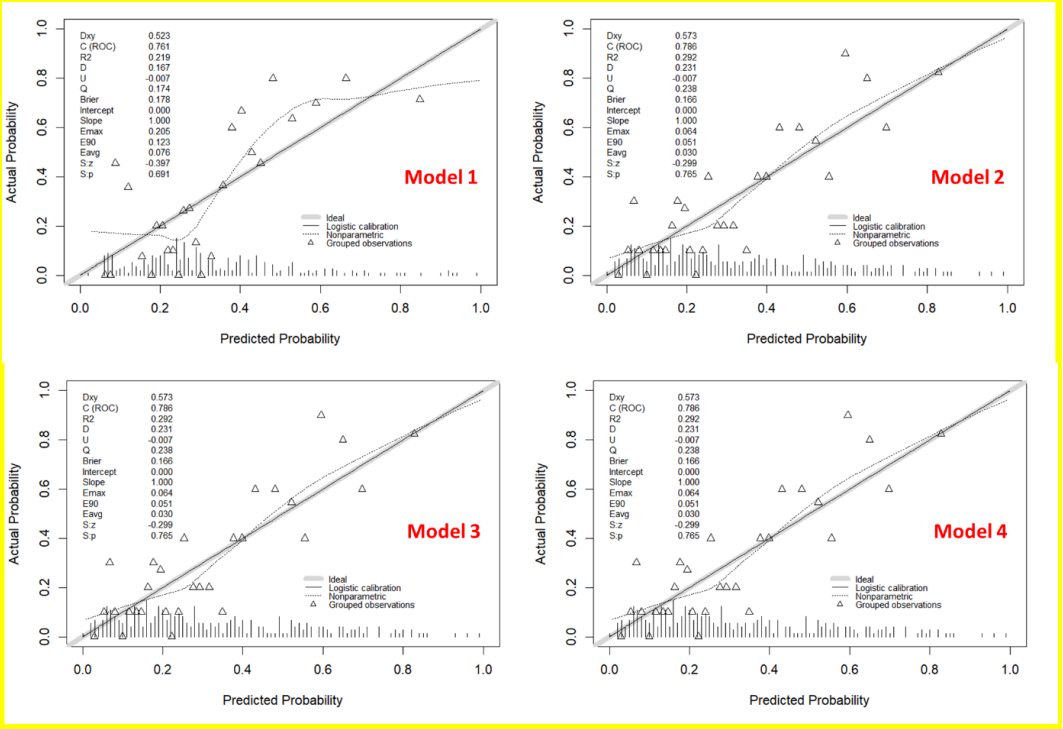

Figure 1. Calibration plots of logistic regression models predicting mortality in heart failure patients. Each plot compares predicted probabilities to observed outcomes using three calibration curves: ideal (diagonal), logistic calibration, and nonparametric smooth. Calibration performance is summarized with metrics including Brier score, calibration slope and intercept, Emax, and Eavg. Calibration-in-the-large (intercept ≈ 0) and slope (≈1) support model reliability for risk prediction in clinical settings.

2.2. Decision curve analysis

DCA was performed to evaluate the clinical utility of the four logistic regression models by estimating their standardized net benefit across a range of high-risk threshold probabilities (**Figure 2**). Net benefit is calculated under the assumption that patients exceeding the risk threshold receive the intervention (e.g., intensified monitoring or treatment), while those below it do not. Model 4 demonstrated the highest net benefit across the broadest threshold range (0.05 to 0.75). However, Model 2 performed comparably well between 0.10 and 0.60, closely overlapping with both Model 3 and Model 4, particularly within the moderate-risk decision range (0.20–0.40). Model 3 added serum sodium to Model 2 but showed negligible improvement in net benefit, aligning with NRI results indicating no added reclassification value. Model 1, which lacked age, underperformed across most thresholds, with utility evident only from 0.10 to 0.45 and occasionally approaching the "treat all" line. From a cost–benefit perspective, a threshold of 0.25 reflects a 3:1 trade-off, where missing a true positive is deemed three times more harmful than treating a false positive. In this context, Model 2 possesses the best balance, retaining strong clinical utility without the added complexity of broader predictor sets.

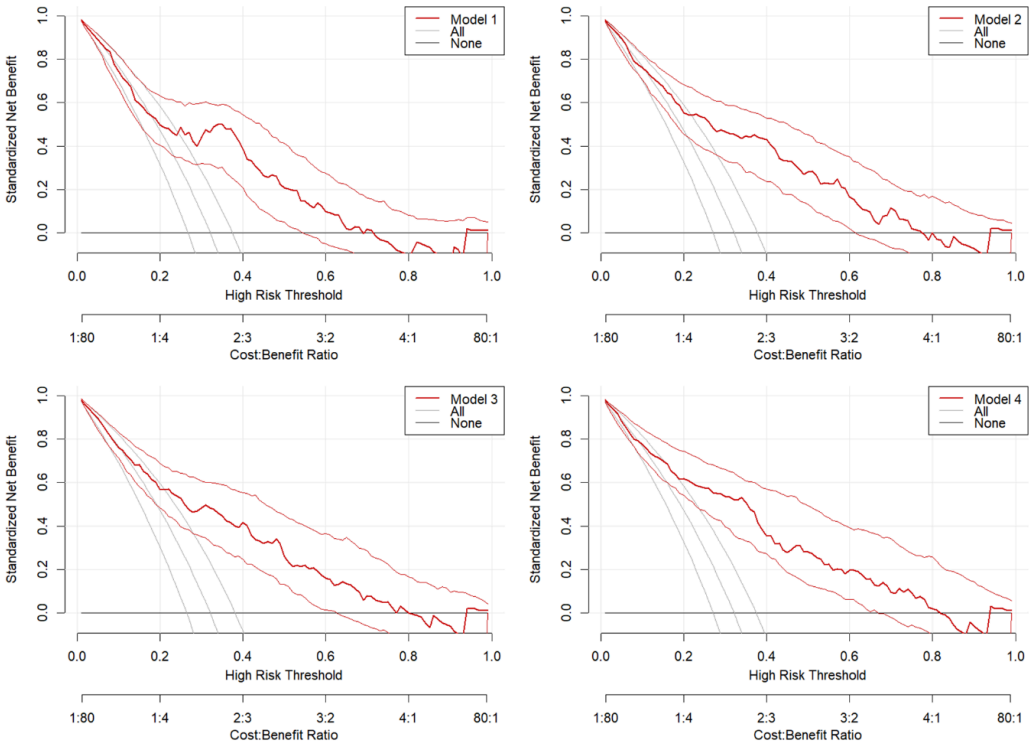

Figure 2. Decision curve analysis comparing clinical utility across logistic regression models. Standardized net benefit is plotted against varying high-risk thresholds for Models 1 to 4. The "All" and "None" strategies are included as references.

2.3. Net Reclassification Improvement

Comparisons across logistic regression models consisting of serum creatinine, ejection fraction, and age consistently matched or outperformed more complex alternatives when evaluated using reclassification metrics, as presented in **Figure 3**. When comparing this model (Model 2) to the baseline two-variable model (Model 1), the overall NRI was 0.111, driven by a positive reclassification among cases (NRI+ = 0.146) and a slight negative reclassification among controls (NRI− = –0.034). Specifically, the model correctly reclassified 19.8% of cases into higher-risk categories (Pr[Up|Case]) and only misclassified 5.2% of them downward (Pr[Down|Case]). For controls, the model incorrectly moved 3.9% into higher risk (Pr[Up|Ctrl]) while correctly reclassified 0.5% downward (Pr[Down|Ctrl]). These values were identical when comparing Model 1 to both Model 3 (which added serum sodium) and Model 4 (the full 11-variable model), indicating that the gains in reclassification were entirely captured by the inclusion of age, and further additions did not enhance patient classification. Strikingly, Model 2 vs Model 3, Model 2 vs Model 4, and Model 3 vs Model 4 all showed NRI = 0, with no patients being reclassified in either direction—an outcome suggesting complete redundancy of the added variables in these contexts. Given its consistently superior performance across both clinical utility and classification metrics, Model 2 was selected for further development into a point-based scoring system.

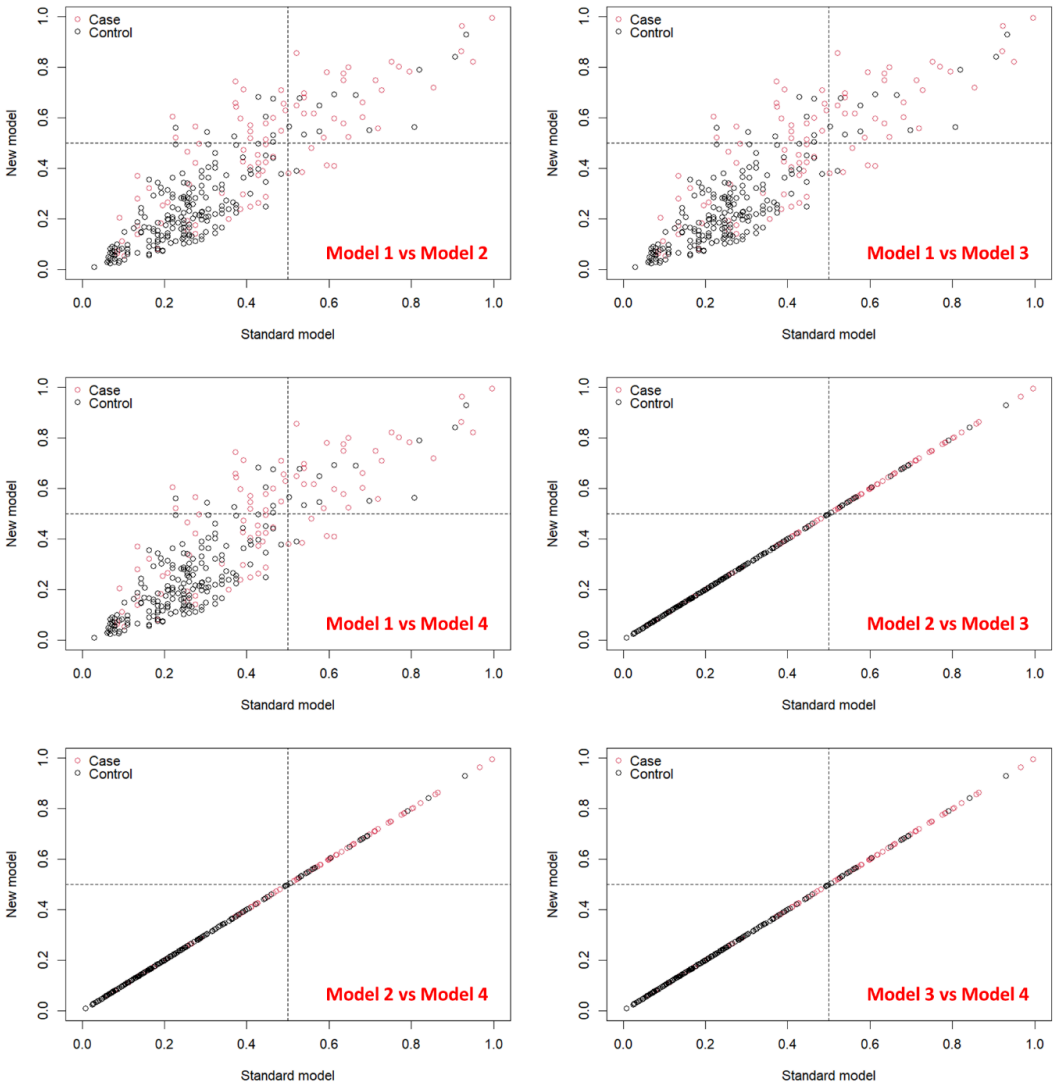

Figure 3. Net Reclassification Improvement (NRI) scatter plots comparing paired logistic regression models. Each panel shows the predicted probabilities from a new model (Y-axis) against a standard model (X-axis) for both cases (red) and controls (black). Dashed lines represent the decision threshold of 0.5. Points above or below the diagonal quadrants represent individuals reclassified by the new model.

2.4. Point-based scoring system

Based on the point-based scoring system, each 1 mg/dL increase in serum creatinine contributes +12 points, each 1% decrease in ejection fraction contributes –1 point, and each 1-year increase in age adds +1 point to the total score. Risk stratification is defined as follows: total scores ≤32 indicate low risk, scores between 33 and 54 indicate moderate risk, and scores ≥55 indicate high risk of mortality. This scoring framework offers a simple and interpretable method for quantifying heart failure mortality risk using three routinely available clinical parameters.

The relationship between total risk score and predicted mortality probability derived from the logistic regression model is presented in **Figure 4**. The curve demonstrates a clear sigmoid pattern, with risk escalating sharply in the mid-range scores. Low-risk patients (score ≤ 32) were associated with predicted risks below approximately 33%, while high-risk individuals (score ≥ 55) approached predicted risks near or exceeding 80%. The moderate-risk category (scores 33–54) captured the inflection zone, with predicted probabilities ranging from roughly 34% to 66%. This visual representation reinforces the scoring system's clinical utility, highlighting its ability to

stratify patients along a continuous risk spectrum with interpretable thresholds for decision-making.

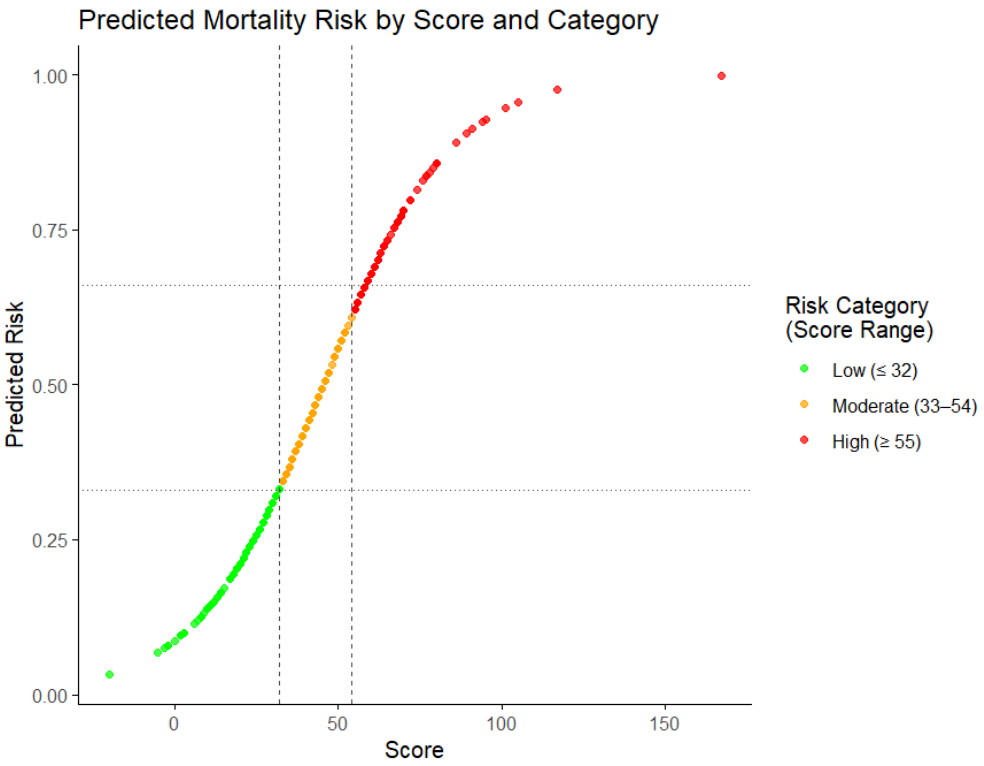

Figure 4. Predicted mortality risk plotted against the total score from the point-based system derived from serum creatinine, ejection fraction, and age. Risk categories are color-coded: green for low risk (score ≤ 32), orange for moderate risk (33–54), and red for high risk (≥ 55). Dashed vertical lines represent category boundaries. The curve shows a sigmoid relationship between score and predicted mortality risk, based on the logistic function.

Table 2. Point-based scoring system derived from the best-performing logistic regression model using serum creatinine, ejection fraction, and age

Scoring value					**Risk stratification**
Predictor	Category / Value Range	Coefficient (β)	Scaled Score	Risk Category	Score Range
Serum Creatinine	Per 1 mg/dL increase	0.83	+12	Low	≤ 32
Ejection Fraction	Per 1% decrease	-0.06	-1	Moderate	33 to 54
Age	Per 1 year increase	0.07	+1	High	≥ 55

The three-category scoring system demonstrated excellent classification performance. Overall accuracy reached 92.3% (95% CI: 88.7%–95.1%), with a Kappa statistic of 0.88, indicating very strong agreement beyond chance. Sensitivity was perfect (1.00) for both the Low- and High-risk categories and remained high for the Moderate group (0.827) (**Table 3**). Specificity was also high, ranging from 0.933 (High) to 1.00 (Moderate), reflecting a low rate of false positives. Precision (positive predictive value)

was perfect for Moderate-risk patients (1.00), high for Low-risk (0.938), and moderately high for High-risk (0.79), the latter potentially affected by a small number of Moderate-risk patients being over-classified. Balanced accuracy across all categories exceeded 0.91, with the Low-risk group achieving the highest value (0.982) (**Table 3**).

Table 3. Performance metrics of the three-category risk classification derived from the point-based scoring system

Class	Sensitivity	Specificity	Precision (PPV)	Balanced accuracy
Low	1	0.964	0.938	0.982
Moderate	0.827	1	1	0.914
High	1	0.933	0.79	0.967

2.5 Case example of the scoring system utility

A 65-year-old patient with heart failure presents with a serum creatinine level of 2.0 mg/dL, an ejection fraction of 30%, and is 65 years old. Using the point-based scoring system, the serum creatinine contributes 24 points (12 points per 1 mg/dL), the reduced ejection fraction contributes –40 points (–1 point for each percentage point below 70%), and age contributes 65 points (1 point per year). The total score is therefore 49 points. According to the model, this score corresponds to an estimated mortality risk of approximately 54.5%, categorizing the patient as moderate risk (score range: 33–54; estimated risk: 34–66%).

Table 4. Illustration of score calculation for a 65-year-old heart failure patient using the point-based scoring system.

Predictor	Patient Value	Scoring Rule	Score Contribution
Serum creatinine (mg/dL)	2.0	12 points per 1 mg/dL	2 × 12 = 24
Ejection fraction (%)	30	–1 point per 1% decrease	(70–30) × –1 = –40
Age (years)	65	+1 point per year	65 × 1 = 65
Total Score			49 points

3. Discussion

Our findings demonstrate that a simple model using only serum creatinine, ejection fraction, and age (Model 2) provides strong predictive performance and clear clinical utility in identifying heart failure patients at high risk of mortality. Despite its parsimony, the model achieved comparable discrimination and net clinical benefit to a full multivariable model while maintaining better calibration and interpretability. From a cost–benefit perspective, the selected predictors are routinely measured in standard heart failure management, requiring no additional testing or financial burden, thus supporting broader implementation in resource-limited settings [13,14]. The use of age—a universally available demographic factor—adds incremental predictive value at no cost. While previous studies have highlighted serum creatinine and ejection fraction as key predictors, few have formally assessed the added value of age within a structured clinical decision framework [12,15–17].

Herein, we found that Model 2 consistently showed the highest net benefit across a wide range of clinically relevant thresholds (0.1–0.6). This finding implies that, for decision thresholds where a clinician would consider initiating treatment or enhanced monitoring if a patient's predicted probability of death exceeds 10% to 60%, Model 2 provides the most favorable balance between true positives and false positives. In

practical terms, using Model 2 to stratify patients during this threshold range would lead to more appropriate interventions for patients who are genuinely at risk of dying, while minimizing unnecessary treatments for those who are not. For instance, at a threshold of 0.2 (i.e., treating those with ≥20% predicted risk), a model with higher net benefit ensures that more actual deaths are correctly identified (true positives), without increasing the number of false alarms (false positives) beyond what is clinically tolerable.

This reinforces the model's clinical utility, as it not only predicts well (discrimination and calibration) but also aligns with decision-making scenarios that matter most in practice. By doing so with only three routinely available variables, Model 2 strikes a favorable balance between effectiveness, simplicity, and feasibility—key elements of real-world cost–benefit trade-offs in heart failure care. To our knowledge, no prior study has integrated these three variables into a logistic model and evaluated its efficiency using a comprehensive suite of validation metrics—including AUC, calibration plots, DCA, and NRI.

The identification of serum creatinine, ejection fraction, and age as the strongest predictors of heart failure mortality reflects both their robust statistical performance and well-established biological plausibility. Serum creatinine, a marker of renal function, is elevated in patients with impaired kidney performance and is a key indicator of cardiorenal syndrome, an intertwined condition commonly seen in heart failure that is associated with poor prognosis [13,15]. Ejection fraction, a direct measure of left ventricular systolic function, is central to diagnosing and managing heart failure with reduced ejection fraction (HFrEF) [18,19]. Age, though non-modifiable, captures the cumulative burden of comorbidities, structural cardiac remodeling, and physiological decline, all of which contribute to worsened outcomes [1,20]. In our present study, a logistic regression model incorporating only these three variables achieved strong predictive performance, with an AUC of 0.786, a favorable Brier score (0.178), and superior calibration and reclassification compared to more complex models, demonstrating their adequacy for simplified yet accurate risk modeling.

These findings hold meaningful clinical implications, particularly in resource-constrained settings. A model based solely on serum creatinine, ejection fraction, and age offers a parsimonious and interpretable approach to early risk stratification. By converting model coefficients into a simple point-based scoring system, clinicians can assign mortality risk levels—low (≤32), moderate (33–54), or high (≥55)—without the need for advanced computation. This facilitates rapid bedside decision-making using only routinely collected variables. The classification system demonstrated strong predictive performance, with high sensitivity and specificity across all risk categories. For instance, the high-risk group achieved a sensitivity of 1.000, specificity of 0.933, and precision (positive predictive value) of 0.790; the low-risk group showed perfect sensitivity (1.000), 0.964 specificity, and 0.938 precision. The balanced accuracy was above 0.96 for both the low and high-risk categories. These metrics support the model's reliability and practical utility in frontline care, especially where access to full diagnostic panels is limited.

Notably, several other variables—such as anaemia, serum sodium, creatinine phosphokinase (CPK), and platelet count—did not significantly enhance prediction metrics when added to the model. While these factors have previously been associated with heart failure outcomes—anaemia through reduced oxygen-carrying capacity [21], hyponatremia through neurohormonal dysregulation [22], elevated CPK as a marker of myocardial injury [23], and platelet abnormalities as indicators of inflammation or thrombogenesis [24], they failed to meaningfully improve discrimination, calibration, or net reclassification in our analysis. Their diminished impact may reflect collinearity, variability in measurement, or context-specific relevance, particularly when core predictors are already accounted for [25].

Our findings, herein, align with and build upon previous studies highlighting the prognostic value of serum creatinine, EF, and age in heart failure mortality prediction. The previous demonstrated that a minimalist model incorporating serum creatinine and ejection fraction alone could achieve strong predictive accuracy using machine learning techniques (AUC ≈ 0.85) [12]. Similarly, other two studies corroborated reinforced the prognostic relevance of these variables by employing advanced ensemble methods, consistently achieving high discrimination performance (accuracy around 90%) [26,27]. More recently, the sarcopenia index (SI)—a ratio of serum creatinine to cystatin C—has emerged as an independent predictor of HFpEF, showing significantly lower values in heart failure patients compared to healthy controls and a strong inverse association with left ventricular EF [28]. Additionally, the age–creatinine–ejection fraction (ACEF) score, originally developed for risk stratification in cardiac surgery, has demonstrated predictive value in patients with myocardial infarction with nonobstructive coronary arteries, with higher tertiles of the score significantly associated with increased risk of major adverse cardiovascular events [29]. It is important to note that, unlike the original ACEF formulation, which yields a continuous risk score requiring numerical input and interpretation [29], our model proposed herein offers a more transparent and customizable point-based approach. Its stepwise scoring makes it easier to understand individual predictor contributions, which may aid clinical decision-making and communication, especially when stratifying patients into discrete risk categories.

Despite being reported in many existing studies, critical components such as calibration and clinical decision-making utility are often overlooked. In a systematic review, it is suggested that while machine learning models typically outperform traditional statistical approaches in discrimination, they frequently lack assessments of calibration and external validity [30]. A report addressed clinical utility through DCA using complex high-dimensional models, demonstrating net benefit gains across clinically relevant thresholds, though at the expense of interpretability and feasibility in resource-limited environments [31]. Similarly, another study utilized DCA to evaluate a complex nomogram predicting ICU heart failure mortality, yet omitted reclassification metrics and calibration assessment [16]. Meanwhile, our study uniquely extends this prior work by rigorously evaluating a streamlined logistic regression model using serum creatinine, EF, and age across multiple dimensions critical for clinical deployment: robust discrimination (AUC = 0.786), good calibration (Brier score = 0.178; Hosmer–Lemeshow $p = 0.675$), meaningful improvement in reclassification metrics (NRI = 0.141), and superior net clinical benefit demonstrated by DCA. This holistic approach to model validation directly addresses gaps identified in the published review [30].

This present study employed comprehensive validation framework addresses significant gaps identified by previous systematic reviews, which have noted that many prediction models for heart failure lack critical assessments of calibration and real-world clinical utility [4,6]. Furthermore, the selection of widely available, routinely measured clinical variables—serum creatinine, ejection fraction, and age—enhances the model's practicality and feasibility, especially in resource-constrained healthcare settings. Nevertheless, our findings must be interpreted within the context of several limitations, such as the small sample size from a single regional dataset. This may constrain statistical power and limit the precision of predictive estimates, potentially impacting the stability of performance metrics in broader populations. Moreover, external validation is still required to assess the reproducibility and generalizability of our model across diverse clinical settings, particularly in populations with different demographic and comorbidity profiles. Additionally, the retrospective nature of the dataset limits causal inference, and potential unmeasured confounders—such as medication use, socioeconomic status, or follow-up adherence—may influence outcomes. Future prospective studies and multicenter validations are essential to confirm these findings and support their clinical implementation.

4. Materials and Methods

4.1. Study Design

This study followed a retrospective analytical design aimed at evaluating the clinical utility and cost–benefit trade-offs of logistic regression models for predicting in-hospital mortality. We began by comparing four models of increasing complexity, focusing first on those using routinely available clinical parameters. Prioritizing clinical deployability, we assessed each model using discrimination, calibration, decision curve analysis, and reclassification metrics. This approach allowed us to identify the simplest model offering the greatest utility. We then developed a point-based scoring system from the best-performing model to support practical application in resource-limited settings, where interpretability and ease of use are critical for early risk stratification.

4.2. Data Source and Study Population

We conducted a secondary analysis using a publicly available dataset published in previous studies [12,32]. The dataset includes clinical records of 299 patients with left ventricular systolic dysfunction and prior heart failure classified as New York Heart Association (NYHA) class III or IV. Data were collected between April and December 2015 at the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad, Punjab, Pakistan. The cohort consisted of 105 women and 194 men, aged between 40 and 95 years, with an average follow-up duration of 130 days. The flow of participants through the study is presented in **Figure 5**.

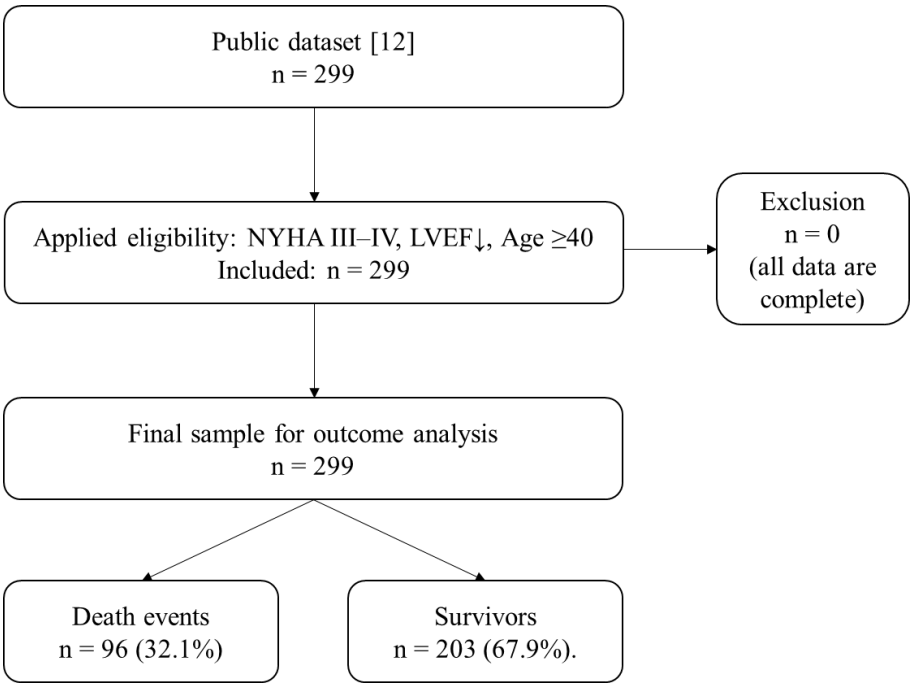

Figure 5. Flow diagram of patient selection and analysis process. All 299 patients with heart failure were included from the original dataset.

4.3. Variables and Outcome

The dataset consisted of 13 variables encompassing demographic, clinical, laboratory, and lifestyle information. Demographic and lifestyle factors included age, sex, and smoking status. Clinical and laboratory variables comprised anemia, diabetes, high blood pressure, platelet count, serum creatinine, serum sodium, creatinine

phosphokinase (CPK), and ejection fraction. Follow-up duration (in days) was also recorded. The primary outcome was all-cause mortality, recorded as a binary variable (1 = death, 0 = survival). The dataset was moderately imbalanced, with 96 death events (32.1%) and 203 survivors (67.9%).

4.4. Model Development

The dataset consisted of 13 variables encompassing demographic, clinical, laboratory, and lifestyle information. Demographic and lifestyle factors included age, sex, and smoking status. Clinical and laboratory variables comprised anemia, diabetes, high blood pressure, platelet count, serum creatinine, serum sodium, creatinine phosphokinase (CPK), and ejection fraction. Follow-up duration (in days) was also recorded. The primary outcome was all-cause mortality, recorded as a binary variable (1 = death, 0 = survival). The dataset was moderately imbalanced, with 96 death events (32.1%) and 203 survivors (67.9%).

4.5. Model Evaluation

4.5.1. Discrimination and Calibration

Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC), which quantifies the ability of each model to distinguish between patients who experienced the event (death) and those who did not. Calibration was assessed using several complementary metrics. The Brier score was calculated to measure overall prediction accuracy, reflecting the mean squared difference between predicted probabilities and actual outcomes. The Hosmer–Lemeshow (HL) goodness-of-fit test was performed by grouping observations into deciles of predicted risk (10 groups), and computing the chi-square statistic to assess alignment between predicted and observed event rates. Additionally, graphical calibration was examined using calibration plots generated, including the visual inspection of the calibration curve as well as extraction of key metrics: calibration slope, intercept, and maximum calibration error (Emax).

4.5.2. Decision Curve Analysis

To evaluate the clinical utility of each predictive model, we conducted DCA that estimates the net benefit of using a predictive model to guide treatment decisions across a continuum of risk threshold probabilities. Net benefit was calculated for each model across threshold probabilities ranging from 0.01 to 0.99 at 0.01 increments. For each threshold, the model's predicted probabilities were used to classify patients as high-risk or not, and the corresponding standardized net benefit was computed. These were then plotted against threshold probabilities to visualize and compare clinical utility. Each model's decision curve was evaluated relative to two default strategies: treating all patients and treating none. These comparisons allow for interpretation of whether a model provides greater clinical benefit than either strategy over clinically relevant threshold ranges. The analysis assumes that intervention (e.g., intensified monitoring or therapy) is only administered to patients classified as high-risk, and that true positives are more beneficial than false positives based on the implied cost:benefit trade-offs at each threshold.

4.5.3. Reclassification Analysis

To evaluate the incremental value of adding predictors across models, we calculated the NRI that quantifies how well a newer model reclassifies individuals into more appropriate risk categories compared to a baseline model. We used a category-based NRI approach with a fixed binary risk threshold of 0.5 to define reclassification into predicted high-risk versus low-risk categories. For each model

comparison, the NRI was decomposed into its components: events correctly reclassified upward, and non-events correctly reclassified downward. To obtain confidence intervals, we performed bootstrap resampling with 1,000 iterations to estimate bias-corrected 95% confidence intervals. All pairwise model comparisons were conducted, including: Model 1 vs Model 2, Model 1 vs Model 3, Model 1 vs Model 4, Model 2 vs Model 3, Model 2 vs Model 4, and Model 3 vs Model 4. This approach enabled formal assessment of whether any additional predictor(s) significantly improved classification performance.

4.6. Methods – Risk Score Derivation and Evaluation

A logistic regression model was developed to predict mortality using serum creatinine, ejection fraction, and age as predictors. Regression coefficients were obtained and scaled relative to the age coefficient to derive integer-based score weights. A total score was calculated for each individual by summing the weighted contributions of all predictors. Risk probabilities were then estimated using the logistic transformation of the total score. Based on these probabilities, individuals were categorized into three predefined risk groups: Low (score ≤ 32 or risk ≤ 33%), Moderate (score 33–54 or risk 34–66%), and High (score ≥ 55 or risk ≥ 67%). To evaluate classification performance, predicted risk categories were compared against observed outcomes, which were grouped to match the three-category structure. Thereafter, a multiclass confusion matrix was generated, and classification performance was assessed through sensitivity, specificity, positive predictive value, and balanced accuracy for each risk group.

4.6.1. Statistical Software

All statistical analyses were performed using R version 4.3.2. Key R packages included stats for logistic regression modeling using the glm() function, pROC for computing the area under the receiver operating characteristic curve (AUC) and associated 95% confidence intervals via bootstrapping, and ResourceSelection for the Hosmer–Lemeshow goodness-of-fit test. Model calibration metrics—including slope, intercept, and maximum calibration error (Emax)—were assessed using the val.prob() function from the rms package. Clinical utility was evaluated through DCA using the rmda package. Reclassification performance was quantified using NRI with the nribin() function from the nricens package, applying 1,000 bootstrap iterations to derive 95% confidence intervals. The caret package was used to generate multiclass confusion matrices and compute classification performance metrics such as accuracy, sensitivity, specificity, precision, and Cohen's kappa. Data management and visualization were supported by tidyverse and ggplot2 packages.

5. Conclusions

A logistic regression model using only serum creatinine, ejection fraction, and age accurately predicts mortality in heart failure patients, achieving strong predictive performance, good calibration, and meaningful clinical utility. This simple, clinically practical model can facilitate timely risk stratification in resource-limited healthcare settings. Future research should validate this model externally in larger and diverse populations, explore its integration into digital health decision-support tools, and assess whether incorporating additional biomarkers or machine learning methods offers meaningful predictive gains without compromising simplicity and interpretability.

6. Patents

Funding: This study received no external funding.

Informed Consent Statement: This study is a secondary analysis of a publicly available, de-identified dataset originally published by Ahmad et al. [32]. The original study was approved by the Institutional Review Board of Government College University, Faisalabad, Pakistan, and was conducted in accordance with the ethical standards outlined in the Declaration of Helsinki (1964) and its subsequent amendments. Written informed consent was obtained from all participants at the time of data collection. As this study involved secondary analysis of anonymized data, no further ethical approval or participant consent was required.

Acknowledgments: Authors appreciate the inter-institutional collaboration during the research and the making of this report.

Conflicts of Interest: The authors declare that they have no known conflicts of interest in relation to the publication of this work.

References

1. Rosch S, Kresoja KP, Besler C, Fengler K, Schoeber A, von Roeder M, et al. Characteristics of heart failure with preserved ejection fraction across the range of left ventricular ejection fraction. Circulation. 2022;146(7):506–18. doi: 10.1161/CIRCULATIONAHA.122.059280.
2. Siddiqi T, Khan Minhas A, Greene S, Van Spall H, Khan S, Pandey A, et al. Trends in heart failure–related mortality among older adults in the United States from 1999-2019. Heart failure. 2022;10(11):851-859. doi: 10.1016/j.jchf.2022.06.012.
3. Gu J, Zheng Z, Li J, Wu S, Sun H, Pang J, et al. Global burden of heart failure in older adults: trends, socioeconomic inequalities, and future projections from 1990 to 2035. European Heart Journal-Quality of Care and Clinical Outcomes. 2025;11(7):1123–36. doi: 10.1093/ehjqcco/qcaf047.
4. Blum M, Gelfman LP, McKendrick K, Pinney SP, Goldstein NE. Enhancing palliative care for patients with advanced heart failure through simple prognostication tools: a comparison of the surprise question, the number of previous heart failure hospitalizations, and the Seattle heart failure model for predicting 1-year survival. Frontiers in Cardiovascular Medicine. 2022;9:836237. doi: 10.3389/fcvm.2022.836237.
5. Akagaki D, Shibata T, Shibao K, Kanaoka K, Nasu T, Ishii S, et al. MAGGIC risk score and drug-related adverse events of sacubitril/valsartan: Insights from the REVIEW-HF registry. IJC Heart & Vasculature. 2025;59:101702. doi: 10.1016/j.ijcha.2025.101702.
6. Gala D, Behl H, Shah M, Makaryus A. The role of artificial intelligence in improving patient outcomes and future of healthcare delivery in cardiology: a narrative review of the literature. Healthcare. 2024;12(4):481 doi: 10.3390/healthcare12040481.
7. Fareed A, Vaid R, Moradeyo A, Sohail A, Sarwar A, Khalid A. Revolutionizing cardiac care: Artificial intelligence applications in heart failure management. Cardiology in Review. 2025;33(5):851. doi: 10.1097/CRD.0000000000000851.
8. Mpanya D, Celik T, Klug E, Ntsinjana H. Machine learning and statistical methods for predicting mortality in heart failure. Heart failure reviews. 2021;26(3):545–52. doi: 10.1007/s10741-020-10052-y.
9. Alabdaljabar MS, Hasan B, Noseworthy PA, Maalouf JF, Ammash NM, Hashmi SK. Machine learning in cardiology: a potential real-world solution in low-and middle-income countries. Journal of multidisciplinary healthcare. 2023;285–95. doi: 10.2147/JMDH.S383810.
10. Razjouyan J, Horstman MJ, Orkaby AR, Virani SS, Intrator O, Goyal P, et al. Developing a parsimonious frailty index for older, multimorbid adults with heart failure using machine learning. The American journal of cardiology. 2023;190:75–81. doi: 10.1016/j.amjcard.2022.11.044.
11. Karatzia L, Aung N, Aksentijevic D. Artificial intelligence in cardiology: Hope for the future and power for the present. Frontiers in Cardiovascular Medicine. 2022;9:945726. doi: 10.3389/fcvm.2022.945726.
12. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC medical informatics and decision making. 2020;20(1):16. doi: 10.1186/s12911-020-1023-5.
13. Fabian J, Vetter B, Bramham K, Luyckx VA, Omosule CL. Point-of-care serum creatinine testing—why, when, where? Oxford University Press; 2024. gfae286 p.
14. Cox ZL, Nandkeolyar S, Johnson AJ, Lindenfeld J, Rali AS. In-hospital initiation and up-titration of guideline-directed medical therapies for heart failure with reduced ejection fraction. Cardiac Failure Review. 2022;8:e21. doi: 10.15420/cfr.2022.08.
15. Li S, Xie X, Zeng X, Wang S, Lan J. Association between serum albumin to serum creatinine ratio and mortality risk in patients with heart failure. Clinical and Translational Science. 2023;16(11):2345–55. doi: 10.1111/cts.13636.
16. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. BMJ open. 2021;11(7):e044779. doi: 10.1136/bmjopen-2020-044779.
17. Verbrugge FH, Omote K, Reddy YN, Sorimachi H, Obokata M, Borlaug BA. Heart failure with preserved ejection fraction in patients with normal natriuretic peptide levels is associated with increased morbidity and mortality. European heart journal. 2022;43(20):1941–51. doi: 10.1093/eurheartj/ehab911.
18. Savarese G, Stolfo D, Sinagra G, Lund LH. Heart failure with mid-range or mildly reduced ejection fraction. Nature Reviews Cardiology. 2022;19(2):100–16. doi: 10.1038/s41569-021-00605-5.
19. Solomonchuk A. Assessment of long-term cardiovascular events in patients with acute myocardial infarction complicated by acute heart failure. Reports of Vinnytsia National Medical University. 2023;27(3):413–8. doi: 10.31393/reports-vnmedical-2023-27(3)-10.
20. Masoli JA, Mensah E, Rajkumar C. Age and ageing cardiovascular collection: blood pressure, coronary heart disease and heart failure. Age and Ageing. 2022;51(8):afac179. doi: 10.1093/ageing/afac179.
21. Sharma YP, Kaur N, Kasinadhuni G, Batta A, Chhabra P, Verma S, et al. Anemia in heart failure: still an unsolved enigma. The Egyptian Heart Journal. 2021;73(1):75. doi: 10.1186/s43044-021-00200-6.
22. Christopoulou E, Liamis G, Naka K, Touloupis P, Gkartzonikas I, Florentin M. Hyponatremia in patients with heart failure beyond the neurohormonal activation associated with reduced cardiac output: A holistic approach. Cardiology. 2022;147(5–6):507–20. doi: 10.1159/000526912.

23. Xue L, Lu W. A Multifactorial approach to explain risk features for predicting survival rate of heart failure. In Springer; 2023. p. 159–72. doi: 10.1007/978-3-031-47126-1_11.
24. Dahlen B, Schulz A, Göbel S, Tröbs SO, Schwuchow-Thonke S, Spronk HM, et al. The impact of platelet indices on clinical outcome in heart failure: results from the MyoVasc study. ESC heart failure. 2021;8(4):2991–3001. doi: 10.1002/ehf2.13390.
25. Leeuwenberg AM, van Smeden M, Langendijk JA, van der Schaaf A, Mauer ME, Moons KG, et al. Performance of binary prediction models in high-correlation low-dimensional settings: a comparison of methods. Diagnostic and prognostic research. 2022;6(1):1. doi: 10.1186/s41512-021-00115-5.
26. Khan NA, Hafiz MFB, Pramanik MA. Enhancing predictive modelling and interpretability in heart failure prediction: a SHAP-based analysis. International Journal of Informatics and Communication Technology (IJ-ICT). 2025;14(1):11. doi: 10.11591/ijict.v14i1.pp11-19.
27. Ahmed S, Shaikh S, Ikram F, Fayaz M, Alwageed HS, Khan F, et al. Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models. Journal of Sensors. 2022;2022(7):1-21. doi: 10.1155/2022/3730303.
28. Wang R, Huang K, Ying H, Duan J, Feng Q, Zhang X, et al. Serum creatinine to cystatin C ratio in relation to heart failure with preserved ejection fraction. BMC Cardiovascular Disorders. 2024;24(1):721. doi: 10.1186/s12872-024-04359-z.
29. Gao S, Ma W, Huang S, Lin X, Yu M. Predictive value of the age, creatinine, and ejection fraction score in patients with myocardial infarction with nonobstructive coronary arteries. Clinical Cardiology. 2021;44(7):1011–8. doi: 10.1002/clc.23650.
30. Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. ESC heart failure. 2021;8(1):106–15. doi: 10.1002/ehf2.13073.
31. Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. Journal of medical Internet research. 2022;24(8):e38082. doi: 10.2196/38082.
32. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: A case study. PloS one. 2017;12(7):e0181001. doi: 10.1371/journal.pone.0181001.