


Article

Development, Validation, and Deployment of a Time-Dependent Machine Learning Model for Predicting One-Year Mortality Risk in Critically Ill Patients with Heart Failure

Jiuyi Wang ^{1,†} , Qingxia Kang ^{2,†}, Shiqi Tian ¹, Shunli Zhang ¹, Kai Wang ^{3,*}  and Guibo Feng ^{1,*}

¹ Department of General Medicine, The Affiliated Yongchuan Hospital of Chongqing Medical University, Chongqing 402160, China; 2023420066@stu.cqmu.edu.cn (J.W.); 2024121535@stu.cqmu.edu.cn (S.T.); zsl2566@163.com (S.Z.)

² Department of Cardiology, The Affiliated Yongchuan Hospital of Chongqing Medical University, Chongqing 402160, China; kqxyhcqmu@163.com

³ Department of Cardiology, The Second Affiliated Hospital of Chongqing Medical University, Chongqing 401336, China

* Correspondence: nkuwangkai@163.com (K.W.); fgbcqmu@163.com (G.F.)

† These authors contributed equally to this work.

Abstract: Background: Heart failure (HF) ranks among the foremost causes of mortality globally, exhibiting particularly high prevalence and significant impact within intensive care units (ICUs). This study sought to develop, validate, and deploy a time-dependent machine learning model aimed at predicting the one-year all-cause mortality risk in ICU patients diagnosed with HF, thereby facilitating precise prognostic evaluation and risk stratification. **Methods:** This study encompassed a cohort of 8960 ICU patients with HF sourced from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database (version 3.1). This latest version of the database added data from 2020 to 2022 on the basis of version 2.2 (covering data from 2008 to 2019); therefore, data spanning 2008 to 2019 ($n = 5748$) were designated for the training set, while data from 2020 to 2022 ($n = 3212$) were reserved for the test set. The primary endpoint of interest was one-year all-cause mortality. Least Absolute Shrinkage and Selection Operator (LASSO) regression was employed to select predictive features from an initial pool of 64 candidate variables (including demographic characteristics, vital signs, comorbidities and complications, therapeutic interventions, routine laboratory data, and disease severity scores). Four predictive models were developed and compared: Cox proportional hazards, random survival forest (RSF), Cox proportional hazards deep neural network (DeepSurv), and eXtreme Gradient Boosting (XGBoost). Model performance was assessed using the concordance index (C-index) and Brier score, with model interpretability addressed through SHapley Additive exPlanations (SHAP) and time-dependent Survival SHapley Additive exPlanations (SurvSHAP(t)). **Results:** This study revealed a one-year mortality rate of 46.1% within the population under investigation. In the training set, LASSO effectively identified 24 features in the model. In the test set, the XGBoost model exhibited superior predictive performance, as evidenced by a C-index of 0.772 and a Brier score of 0.161, outperforming the Cox model (C-index: 0.740, Brier score: 0.175), the RSF model (C-index: 0.747, Brier score: 0.178), and the DeepSurv model (C-index: 0.723, Brier score: 0.183). Decision curve analysis validated the clinical utility of the XGBoost model across a broad spectrum of risk thresholds. Feature importance analysis identified the red cell distribution width-to-albumin ratio (RAR), Charlson Comorbidity Index, Simplified Acute Physiology Score II (SAPS II), Acute Physiology Score III (APS III), and the age–bilirubin–INR–creatinine (ABIC) score as the top five predictive factors. Consequently, an online risk prediction tool based on this model has been developed and is publicly accessible. **Conclusions:** The time-dependent XGBoost model demonstrated



Academic Editors: Hwa Liang Leo, Vida Abedi and Alireza Vafaei Sadr

Received: 9 March 2025

Revised: 14 April 2025

Accepted: 7 May 2025

Published: 12 May 2025

Citation: Wang, J.; Kang, Q.; Tian, S.; Zhang, S.; Wang, K.; Feng, G. Development, Validation, and Deployment of a Time-Dependent Machine Learning Model for Predicting One-Year Mortality Risk in Critically Ill Patients with Heart Failure. *Bioengineering* **2025**, *12*, 511. <https://doi.org/10.3390/bioengineering12050511>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

robust predictive capability in evaluating the one-year all-cause mortality risk in critically ill HF patients. This model offered a useful tool for early risk identification and supported timely interventions.

Keywords: heart failure; intensive care unit; machine learning; time-dependent; random survival forest; XGBoost; MIMIC-IV database

1. Introduction

Heart failure (HF) remains a major and global health issue, marked by persistently high morbidity and mortality rates [1]. From 1990 to 2019, the global prevalence of HF increased by 106.3%, reaching 56.2 million cases [2]. The complex interplay of comorbidities, HF severity, and multi-organ dysfunction necessitates advanced care for critically ill heart failure patients in the intensive care unit (ICU), presenting substantial challenges for comprehensive management. Acute HF signifies a critical juncture in disease progression, as ICU survivors experience a one-year mortality rate of 46.5%, a rate comparable to or higher than that observed in many malignancies [3,4]. Accurate prediction of mortality risk in these patients is essential for guiding clinical decision-making, optimizing resource allocation, and tailoring therapeutic strategies. Although established prognostic tools such as the Seattle Heart Failure Model (SHFM) [5], Get With The Guidelines-Heart Failure (GWTG-HF) [6], and AHEAD (A: atrial fibrillation; H: hemoglobin; E: elderly; A: abnormal renal parameters; and D: diabetes mellitus) score [7] are widely used, their reliance on linear statistical assumptions limits their capacity to capture the complex interactions among clinical variables, thereby constraining predictive accuracy in heterogeneous ICU populations.

Recent advancements in machine learning (ML) present transformative potential for prognostic modeling by excelling in the processing of high-dimensional data and the identification of nonlinear relationships. Firstly, ensemble algorithms (such as eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM)) effectively decode nonlinear biomarker interactions; for instance, XGBoost has demonstrated superior performance compared to logistic regression in distinguishing one-year mortality [8]. Secondly, interpretability frameworks address the challenges posed by black box models; most physicians preferred ML outputs that were augmented with model-agnostic interpretability methods, with significant correlations observed between clinician comprehension and interpretability, as well as between interpretability and trust [9]. Although ML has shown superior performance over traditional methods in predicting heart failure outcomes, critical gaps remain. Specifically, current ML applications often treat mortality as a binary outcome, neglecting the time-dependent survival information [10–12]. Concurrently, validated composite laboratory indices—such as albumin-corrected anion gap (ACAG) [13] and albumin–bilirubin index (ALBI) [14], and red cell distribution width-to-albumin ratio (RAR) [15]—were underutilized in ML frameworks, despite their established prognostic value.

To address these gaps, we developed and validated a time-dependent ML model to predict one-year all-cause mortality in ICU-admitted heart failure patients, utilizing the Medical Information Mart for Intensive Care IV database (MIMIC-IV version 3.1, <https://physionet.org/content/mimiciv/3.1/>, accessed on 24 November 2024) [16]. Our approach uniquely integrates routine clinical parameters with validated prognostic composites while employing advanced techniques for model interpretation, validation, and online deployment. We developed the first time-dependent machine learning model specifically for critically ill heart failure patients in the ICU setting that is capable of providing dynamic risk predictions across the 365-day post-discharge timeline. Our approach integrates novel

composite indices (such as RAR and ABIC) with traditional prognostic scores, enhancing predictive accuracy beyond what either approach can achieve independently. By implementing advanced interpretability techniques (SurvSHAP(t)), we provide time-dependent feature importance analysis, offering insights into how predictors' influence changes over the follow-up period. To translate these complex models into clinical practice, we deployed a free, web-based calculator that serves as an accessible decision support tool for bedside use. Notably, our model demonstrates generalizability across the COVID-19 pandemic period, suggesting robust performance despite significant healthcare disruption. Our key findings reveal that the XGBoost-based model achieved superior performance compared to traditional approaches, with novel laboratory indices and composite scores emerging as the most influential predictors.

The remainder of this manuscript is organized as follows: In Section 2, we describe our study population, data collection procedures, feature extraction, and the development of four different machine learning models. Section 3 presents the baseline characteristics of our study cohort, the performance metrics of the various models, feature importance analysis, and details of our web-based calculator deployment. In Section 4, we contextualize our findings within the existing literature, explore the clinical implications of our model, acknowledge limitations, and suggest directions for future research. Finally, Section 5 summarizes the key findings and potential impact of our work. The Supplementary Materials provide additional methodological details and supplementary analyses.

2. Materials and Methods

2.1. Research Problem and Objectives

This study addresses the critical research problem of accurately predicting one-year mortality risk in critically ill heart failure patients following ICU admission. Specifically, we aimed to develop a time-dependent machine learning model that can (1) predict one-year all-cause mortality risk dynamically, enabling survival probability estimation at any time point within 365 days post-discharge; (2) integrate composite biomarkers (e.g., RAR, ABIC) with routine clinical parameters to enhance predictive accuracy and pathophysiological relevance; and (3) validate the model's clinical utility through rigorous temporal calibration, interpretability frameworks (SHAP/SurvSHAP(t)), and deployment as an open-access tool.

2.2. Sample Size and Study Population

This retrospective cohort study utilized the MIMIC-IV database (version 3.1), a comprehensive and publicly accessible critical care repository containing detailed clinical data from critically ill patients treated at Beth Israel Deaconess Medical Center between 2008 and 2022. Study participants were selected based on predefined inclusion and exclusion criteria, with inclusion requiring a definitive diagnosis of congestive HF (International Classification of Diseases, 9th Revision [ICD-9] code 428* and 10th Revision [ICD-10] code I50*). Exclusion criteria comprised (1) age < 18 years; (2) no prior ICU admission history; (3) ICU length of stay < 24 h; and (4) missing albumin measurements. The patient selection flowchart outlined the screening process, resulting in a final analytical cohort of 8960 patients (Figure 1).

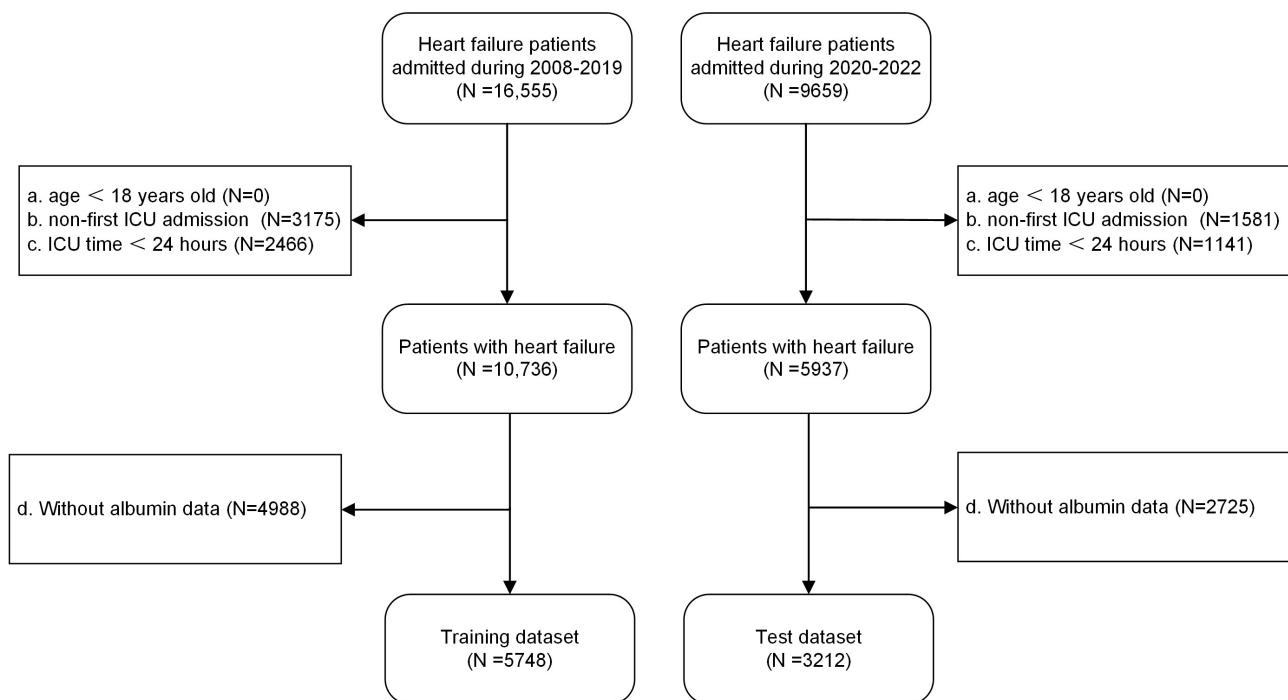


Figure 1. Flow chart of study population inclusion.

This study strictly adhered to ethical guidelines established by institutional review boards, national research committees, and the 1964 Declaration of Helsinki and its subsequent amendments. Informed consent was waived as the research involved secondary analysis of de-identified public data without direct or indirect identifiers. Furthermore, no additional ethical review was required given the use of publicly available datasets.

2.3. Data Collection and Outcome Definition

Within 24 h of ICU admission, the initial measurement was used to extract clinical data. Data points with over 20% missing values were excluded. To tackle the problem of incomplete data, the multiple imputation technique was applied based on the missing-at-random (MAR) assumption. In particular, variables with less than 20% missing information underwent multiple imputation through chained equations (MICE) with the help of the mice package, generating five complete datasets over 50 iterations, after which Rubin's rules were applied for pooled estimation.

Clinical data were systematically extracted according to a predefined protocol, encompassing demographic characteristics (age, sex), vital signs (heart rate, respiratory rate, oxygen saturation (SpO₂), body temperature, systolic and diastolic blood pressure), comorbidities and complications (diabetes mellitus, hypertension, chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), acute kidney injury (AKI) and AKI staging, and sepsis), and cardiac function parameter (left ventricular ejection fraction (LVEF)). Therapeutic interventions were categorized as continuous renal replacement therapy (CRRT) or mechanical ventilation. Disease severity was evaluated using multiple scoring systems, including Sequential Organ Failure Assessment (SOFA), Acute Physiology Score III (APS III), Acute Physiology and Chronic Health Evaluation (APACHE) III, Glasgow Coma Scale (GCS), Simplified Acute Physiology Score II (SAPS II), Charlson Comorbidity Index (CCI), Systemic Inflammatory Response Syndrome (SIRS) score, and Oxford Acute Severity of Illness Score (OASIS).

Laboratory parameters included complete blood count indices (white blood cell count (WBC), red blood cell count (RBC), hemoglobin, hematocrit, platelet count, erythrocyte

indices (mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and red cell distribution width (RDW)), leukocyte differentials (neutrophils, lymphocytes, basophils, and eosinophils), liver function tests (alanine aminotransferase (ALT), aspartate aminotransferase (AST), albumin, and total bilirubin), renal function markers (creatinine, blood urea nitrogen (BUN)), serum glucose, electrolyte and acid-base balance parameters (sodium, calcium, chloride, and potassium), coagulation profiles (international normalized ratio (INR), prothrombin time (PT), activated partial thromboplastin time (PTT)), and arterial blood gas measurements (pH, partial pressure of carbon dioxide (PCO₂), partial pressure of oxygen (PO₂), lactate, and anion gap).

In alignment with published prognostic evidence, validated composite indices were calculated, including RAR, hemoglobin-to-RDW ratio (HRR), BUN-to-creatinine ratio (BCR), BUN-to-albumin ratio (BAR), albumin-to-creatinine ratio (ACR), creatinine-to-total bilirubin ratio, ALT-to-AST ratio, ACAG, ALBI, lactate-to-albumin ratio, sepsis-induced coagulopathy (SIC) score, and the age–bilirubin–INR–creatinine (ABIC) score.

The primary endpoint was defined as all-cause mortality within one year post-discharge. All parameters, including their scales (integer or interval values) and measurement units, were fully detailed in Table S1.

2.4. Statistical Analysis

2.4.1. Data Preprocessing

The analytical matrix was developed from an initial set of 64 clinical predictor variables, with standardized preprocessing procedures implemented to ensure data quality. Continuous variables were normalized to a range of [0, 1] using min–max scaling to mitigate the impact of scale differences on ML algorithms. Categorical variables were converted into orthogonal dummy variables through one-hot encoding. Under the assumption of missing-at-random (MAR) data, variables with ≤20% missingness underwent multiple imputation via chained equations (MICE) using the mice package, generating five complete datasets through 50 iterations, followed by Rubin’s rule for pooled estimates [17].

Continuous variables were reported as medians with interquartile ranges, and between-group comparisons were conducted using Mann–Whitney U tests. Categorical variables were presented as counts and percentages, with analyses performed using chi-square tests. The two-tailed *p* value < 0.05 was considered statistically significant. All statistical analyses were conducted using R version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>, accessed on 17 September 2022).

2.4.2. Model Development and Evaluation

Sample size estimation was guided by the four-step predictive model methodology proposed by Riley et al. [18], and the estimated sample size was compared against the actual cohort size. Patients admitted between 2008 and 2019 were assigned to the training set, while those admitted from 2020 to 2022 constituted the temporally isolated test set. This chronological partitioning was employed to maintain the independence of the test set, thereby reducing the risk of overfitting and enhancing calibration stability during external validation.

In the training cohort, global feature selection was achieved using five-fold cross-validated Least Absolute Shrinkage and Selection Operator (LASSO) regression. Four ML models—Cox proportional hazards regression, random survival forest (RSF), eXtreme Gradient Boosting (XGBoost), and Cox proportional hazards deep neural network (DeepSurv)—were developed utilizing features selected through LASSO to predict the

study endpoint. Hyperparameter optimization was conducted using five-fold cross-validation combined with Bayesian optimization.

In the test cohort, model discrimination was evaluated using Harrell's concordance index (C-index) and the integrated cumulative/dynamic area under the receiver operating characteristic curve (C/D AUC). Model calibration was assessed via Brier scores and calibration plots, while clinical utility was examined through decision curve analysis (DCA) and clinical impact curve (CIC). The optimal model was determined based on the highest C-index and the lowest Brier score.

2.4.3. Model Interpretation

To address the inherent "black box" nature of ML models, interpretability frameworks were systematically applied. Global explanations were provided using SHapley Additive exPlanations (SHAP) analysis, time-dependent variable importance metrics, and partial dependence plots.

Local interpretability for individual predictions was achieved through SurvLIME and time-dependent Survival SHapley Additive Explanations (SurvSHAP(t)), which quantified feature contributions to risk predictions across both temporal and population dimensions.

2.4.4. Model Deployment

To advance clinical translation, an online risk prediction calculator was developed utilizing the R Shiny framework. This tool integrates an optimized model, allowing for the real-time generation of individualized mortality risk predictions based on clinician-provided parameters, thereby offering intuitive, data-driven support for clinical decision-making.

3. Results

3.1. Sample Size and Baseline Characteristics

The minimum required sample size was calculated to be 1922 under predefined parameters (Cox–Snell R-squared = 0.26, event rate = 45.32%, mean follow-up = 0.63 years, and 64 candidate predictors), ensuring sufficient statistical power. This study ultimately included 5748 HF patients admitted between 2008 and 2019, and 3212 patients from 2020 to 2022, thereby exceeding the minimum requirement by more than threefold (Supplementary Materials, Table S1).

Baseline characteristics indicated comparable demographics between the datasets. In the training and test sets, median age was 72.54 years and 72.42 years, with female predominance of 56.3% and 57.2%. Prevalence of comorbidities and complications included diabetes (21.3% and 19.5%), hypertension (37.2% and 34.1%), CKD (23.7% and 21.3%), and AKI (79.9% and 63.9%). Mechanical ventilation was required in 87.5% and 75.7%. The primary endpoint (one-year mortality) occurred in 45.3% (2605/5748) of the training set and 47.6% (1530/3212) of the test set, with no significant temporal discrepancies in event rates ($p = 0.120$).

3.2. Feature Selection and Model Development

In the training set, LASSO regression was employed for automated feature selection. By adjusting the regularization coefficient lambda (λ) to minimize the loss function (binomial deviance), LASSO regression produced sparse coefficients, ultimately selecting 24 out of 64 features for inclusion in the ML models. These features were identified at an optimal shrinkage parameter (λ_{1se}) of 0.0276 (Figure 2).

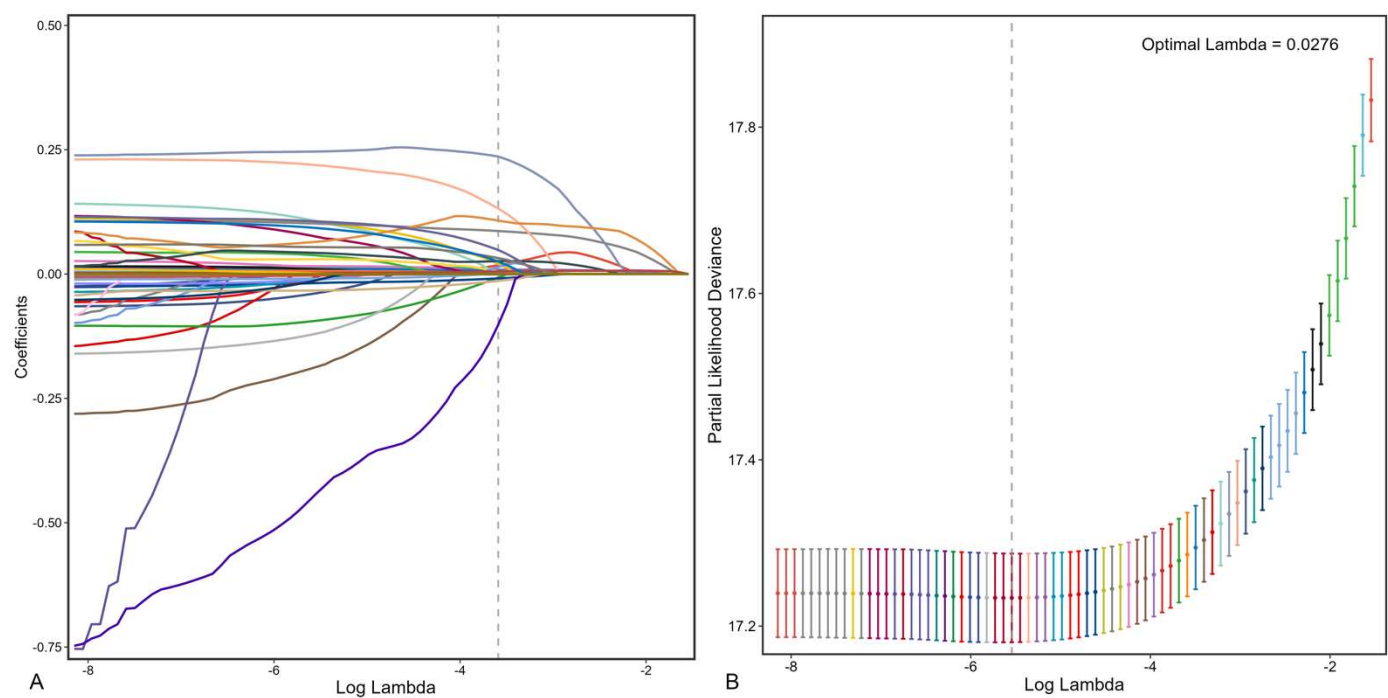


Figure 2. Feature selection process. Automated feature selection for 64 clinical factors was performed using Least Absolute Shrinkage and Selection Operator, which minimized the loss function binomial deviance, shrank coefficients, and produced some coefficients that are zero, allowing efficient feature selection (A). The algorithm outputted 24 filtered features with non-zero coefficients that were included in model generation subsequently (B).

Four ML models were developed utilizing these 24 features. Through the five-fold cross-validation, the optimized hyperparameters for each model are detailed in Table 1. Following hyperparameter optimization, the models were retrained on the complete training dataset.

Table 1. Optimal parameters of machine learning models in predicting one-year mortality.

Models	Optimal Parameters		
XGBoost	nrounds = 368, nthread = 1	subsample = 0.5488236	
	eta = 0.004817722	colsample_bytree = 0.5026403	
	max_depth = 8	lambda = 0.1330041	
RSF	min_child_weight = 4.4415141	alpha = 2.646525	
	num.trees = 264, mtry = 2	num.threads = 1	
	min.node.size = 5	max.depth = 10	
DeepSur	num_nodes = 246	batch_norm = TRUE	
	learning_rate = 0.00111408	activation = "sigmoid"	
	dropout = 0.3304397	optimizer = "adamax"	

3.3. Model Evaluation

A comprehensive evaluation was conducted on the test cohort using metrics including the C-index, Brier score, recall rate, and D-calibration (Supplementary Materials, Figure S1). The XGBoost model exhibited superior discriminative performance, achieving a C-index of 0.772, surpassing DeepSurv (0.714), Cox regression (0.740), and RSF (0.748). XGBoost consistently outperformed the other models across secondary metrics (Figure 3A) and throughout the follow-up period (Figure 3B).



Figure 3. Model performance for the whole cohort. Explainable machine learning (XAI) data are shown as bar plots (A). Explainable machine learning (XAI) was used as a time-dependent estimation (B).

Calibration analysis indicated a strong concordance between predicted and observed mortality probabilities across all four models (Supplementary Materials, Figure S2). The XGBoost model demonstrated superior calibration performance, with a Brier score of 0.165, compared to DeepSurv (0.190), Cox regression (0.175), and RSF (0.177).

In addition, DCA indicated that the XGBoost model provided a greater net benefit compared to both the “zero mortality risk” and “all mortality” strategies, surpassing the performance of the other three models across threshold probabilities ranging from 30% to 100% (Supplementary Materials, Figure S3A). The CIC analysis further demonstrated that within this threshold range, the XGBoost model significantly minimized unnecessary interventions while enhancing risk-avoidance efficacy, thereby exhibiting superior overall

intervention efficiency relative to other models (Supplementary Materials, Figure S3B). Consequently, following a comprehensive evaluation of all metrics, the XGBoost model was identified as the optimal model.

3.4. Model Interpretation and Online Deployment

3.4.1. Global Explanations

SHAP analysis was utilized to rank feature importance within the XGBoost model (Figure 4A), with RAR, CCI, SAPS II, APS III, and ABIC emerging as the top five predictors. Time-dependent permutation importance analysis identified RAR, ABIC, and BCR as the most influential factors for overall survival (Figure 4B). Partial dependence plots revealed nonlinear relationships between predictors and survival, with RAR showing the most pronounced negative impact on overall survival (Supplementary Materials, Figure S1).

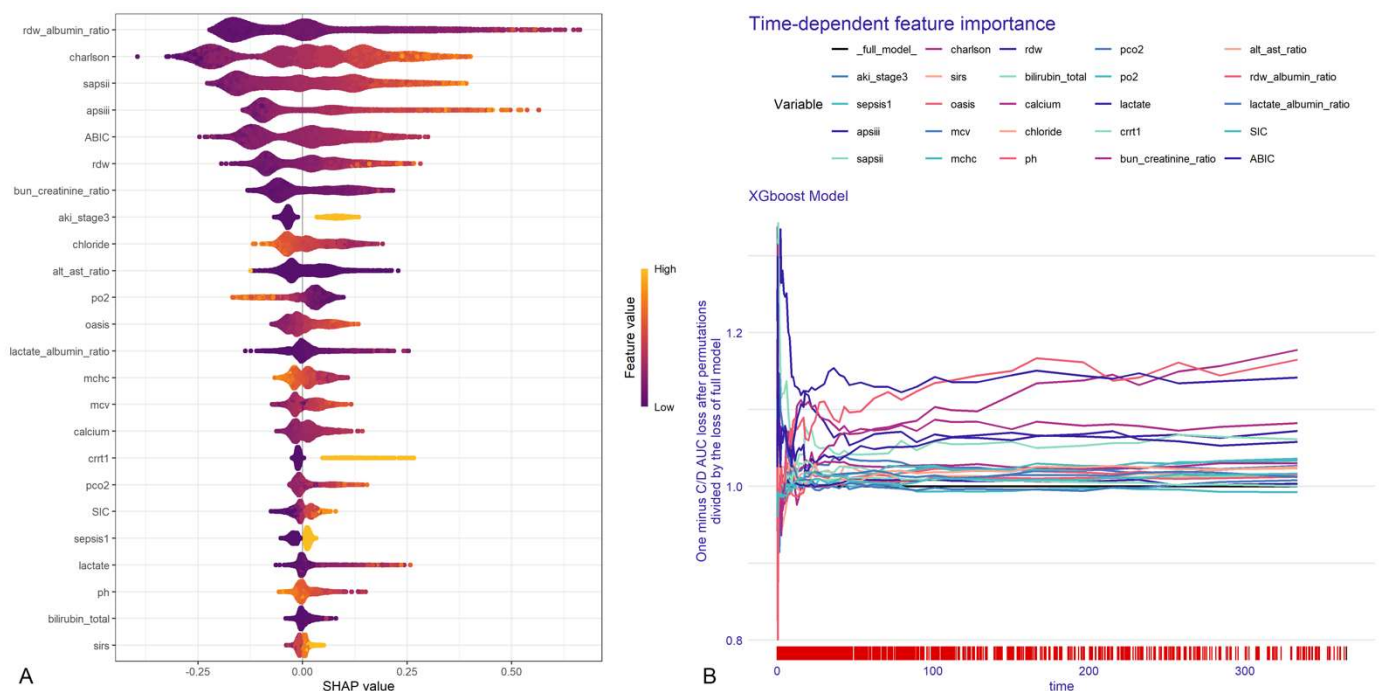


Figure 4. Chart of feature importance ranking in test data (A). The importance ranking of the 24 risk factors with stability and interpretation using the optimal model. Each point in the graph represents the SHAP value for each sample; a color closer to purple indicates a larger value, while one closer to yellow indicates a smaller value. The more scattered the points in the graph, the greater the influence of the variable on the model. Time-dependent feature importance for the whole cohort and C/D AUC loss after permutation (B).

3.4.2. Local Explanations

The SurvSHAP(t) algorithm was applied to examine the time-dependent survival contributions of individual risk factors for selected patients. The y-axis represents SurvSHAP(t) values, where positive values indicate variables that increase overall survival probability, and negative values denote variables that decrease overall survival probability. Patient 1593, diagnosed with heart failure, was randomly selected from the survival cohort to illustrate the translation from population-level predictions to individual-level predictions. The analysis revealed that the variables APS III and SAPS II enhanced this patient's survival probability, whereas RAR and BCR reduced it (Supplementary Materials, Figure S4C).

Although SurvLIME plots were conceptually similar to SurvSHAP(t), they provided distinct mechanistic insights into individualized risk factors. The left panel quantified the impact of covariates on survival probability, with larger shaded areas indicating stronger

effects and higher local importance values correlating with decreased survival likelihood. The right panel compares model-predicted survival curves with SurvLIME-derived approximations; a closer alignment between the two functions validated the fidelity of the explanation. Upon reanalysis of Patient 1593 using XGBoost, SurvLIME identified serum chloride (detrimental) and pH (protective) as critical modifiers of prognosis (Supplementary Materials, Figure S4A). The strong concordance between SurvLIME-explained and model-predicted survival trajectories confirmed the accuracy of individualized risk estimation (Supplementary Materials, Figure S4B).

3.4.3. Model Deployment

The web-based implementation of the XGBoost model was universally accessible to clinicians at <https://cqmuwjy-app-for-mortality-prediction-app.shinyapps.io/deployment-1/>, accessed on 25 January 2025. By inputting values for the 24 predefined clinical parameters, this tool automatically generated survival probability predictions for patients with HF admitted to the ICU at any time point within 365 days post-discharge (Figure 5).

Machine learning app for
1-year mortality prediction

AKI Stage 3
1

Sepsis
0

APACHE II
40

SAPS II
50

Charlson
3

SIRS
2

OASIS
30

MCV
90

MCHC
33

RDW
14

Total Bilirubin
1.5

Calcium
9

Chloride
100

pH
7.4

pCO2
40

pO2
80

Lactate
1.5

CRR
1

BUN/Creatinine Ratio
20

ALT/AST Ratio
1.7

RDW/Albumin Ratio
8.8

Lactate/Albumin Ratio
4

SIC
3

ABIC
8

time-point(days)
200

Predict

Survival probability:
survival probability at 200 days is: 0.5691"

Figure 5. Online deployment interface: ICU congestive heart failure patient mortality risk prediction platform. Input patient clinical information and click the “Predict” button to obtain real-time, individualized in-hospital mortality risk assessment.

4. Discussion

This study developed and validated an interpretable, time-dependent machine learning model based on the XGBoost algorithm to predict one-year all-cause mortality in ICU patients with HF. The model demonstrated superior predictive performance, with a C-index of 0.772, calibration as indicated by a Brier score of 0.165, and clinical net benefit across a threshold probability range of 30 to 100%. It was deployed online as a personalized risk prediction tool.

The SHFM and AHEAD scores, both used to predict mortality in heart failure patients, have shown suboptimal performance [5]. Similarly, the GWTG-HF model, when applied to assess risk in ICU patients with heart failure, yielded an AUC of 0.649 [19]. Consequently, existing risk scoring systems are neither specifically tailored to the ICU heart failure population nor exhibit outstanding prognostic performance. In contrast, ML has markedly improved predictive accuracy due to its ability to model nonlinear relationships and overcome multicollinearity. Adler et al., utilizing data from 5822 HF patients, demonstrated that ML algorithms enhanced predictive accuracy by 18 to 39% (AUC 0.88) compared to traditional risk scores (MAGGIC/AUC 0.74; ADHERE/AUC 0.63), with a 2.1-fold increase in sensitivity for identifying high-risk patients [20]. A systematic review by Shin et al., encompassing 686,842 patients, further suggested that in most studies focused on predicting readmission and mortality risk in HF patients, ML algorithms possessed superior discriminative capability relative to conventional statistical models [21]. In alignment with these findings, we assessed the long-term performance of four ML models (XGBoost, RSF, DeepSurv, and Cox regression) using 24 clinical features to predict mortality. We observed that the C-index of the XGBoost model was higher than that of the Cox regression model. This advantage could be attributed to several mechanisms: tree-based ensemble models; iteratively correct residuals through a boosting framework, thereby effectively capturing nonlinear interactions among high-dimensional clinical variables; and the integration of regularization techniques, such as L1 and L2 penalty terms (LASSO regression), which effectively addresses the multicollinearity challenges commonly encountered in traditional statistical models. In contrast to deep learning methodologies, which typically necessitate large sample sizes, XGBoost leverages parallel computation and pruning optimization to manage model complexity and mitigate overfitting risks within moderately sized medical datasets ($n = 8960$). This characteristic was corroborated in our study, where the DeepSurv model, despite its greater number of parameters, exhibited markedly inferior performance on the test set compared to XGBoost, indicating a tendency towards overfitting. The performance evaluation conducted on a relatively independent test set over a specified time span offers credible support for assessing the model's generalization capabilities. Moreover, the deployment of the model within a network framework provided a distinct advantage for clinical translation and implementation, surpassing the capabilities of most existing studies.

Presently, the majority of machine learning studies simplify mortality to a binary endpoint, thereby forfeiting valuable temporal information. For instance, the study conducted by Tong et al. demonstrated that RSF and gradient boosting models, which utilized time-dependent analysis, significantly surpassed traditional binary classification models in predicting risk among heart failure patients [22]. These models exhibited superior short-term calibration and efficient utilization of variables, alongside a notable enhancement in the dynamic C-index. This research introduced an innovative application of time-dependent ML, achieving for the first time a continuous prediction of mortality risk in ICU patients with HF. In comparison to the in-hospital mortality model developed by Li et al. [23], our model represented a significant advancement in temporal resolution and clinical applicability. It offered a risk prediction platform capable of providing survival

probabilities at any time point within a 365-day period, thereby offering a quantitative basis for dynamically adjusting treatment strategies.

This model significantly optimized feature engineering by incorporating multidimensional composite indices. Previous studies have suggested that an elevated RAR was associated with systemic inflammatory response, oxidative stress, and increased mortality risk [24]. The ABIC score, which incorporates liver and kidney function alongside coagulation abnormalities, has been validated in patients with coronary heart disease [25]. Our previous research demonstrated that ACAG was strongly positively correlated with mortality in HF and could enhance the predictive value of the SOFA and APS III scores [26]. The ALBI score was found to be independently associated with mortality in these HF patients, exhibiting an even more pronounced prognostic effect in younger patients and those with lower creatinine levels [27]. Additionally, a negative nonlinear relationship exists between ACR and mortality, which was challenging to adequately characterize using traditional Cox models [28]. These findings suggest that composite indices, by reflecting pathophysiological processes through a multi-organ interaction network, possess inherent nonlinear features and high-dimensional correlations that present opportunities for optimized ML. In the present study, we integrated six traditional prognostic scores (e.g., APS III, SAPS II) and eight composite laboratory indices (e.g., RAR, ABIC, ACAG). Notably, the contribution weights of RAR and ABIC in the model's feature importance ranking surpassed those of the traditional scoring systems, indicating that composite laboratory indices—owing to their pathophysiological associations with organ dysfunction and systemic inflammatory response—may offer a more sensitive prognostic signal, thereby complementing previous studies.

The temporal validation set (2020–2022) encompassed the COVID-19 pandemic, a period marked by significant shifts in ICU practices and patient acuity. Despite these systemic disruptions, our model demonstrated stable performance (C-index = 0.772), suggesting that its dependence on composite biomarkers capturing systemic pathophysiology (e.g., RAR, ABIC) may buffer against transient clinical variability. However, the absence of explicit COVID-19 status annotations in the dataset precludes direct analysis of pandemic-specific effects on heart failure outcomes. Future work should investigate model performance in cohorts with confirmed SARS-CoV-2 co-infections to further validate its applicability in pandemic-influenced critical care settings.

Nonetheless, this study has certain limitations. First, the dataset contained missing values; however, multiple imputation methods were utilized to address these gaps, potentially approximating the true values. Second, determining whether HF was the primary cause of ICU admission and identifying the specific cause of patient mortality within the MIMIC database presented significant challenges. ICU admissions frequently resulted from critical conditions that impact multiple organ systems, including HF, which often led to multi-organ failure. From the patient's perspective, all-cause mortality may serve as a more meaningful endpoint. Nonetheless, it was essential to acknowledge the distinct pathophysiological mechanisms that differentiate acute from chronic heart failure, as well as heart failure with reduced versus preserved left ventricular ejection fraction.

5. Conclusions

This study employed time-dependent ML techniques to develop an innovative risk stratification model for categorizing the one-year all-cause mortality risk among ICU patients with HF. The model is accessible through a freely available web-based calculator. This tool not only quantifies mortality risk but also unveils actionable pathophysiological insights. For instance, elevated RAR and ABIC scores may signal systemic inflammation and multi-organ dysfunction, urging clinicians to tailor therapies targeting these pathways

(e.g., albumin supplementation, anti-inflammatory agents). Furthermore, the integration of SHAP/SurvSHAP(t) framework bridges ML interpretability with clinical reasoning, as demonstrated in recent bioinformatics advancements [29,30].

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/bioengineering12050511/s1>, Supplementary Figure S1. The performance metrics of all models in the test dataset. Supplementary Figure S2. The calibration plots at the 365th day for all models in the test dataset. Calibration plots of predicted probabilities (X-axis) and actual proportions (Y-axis) for different prediction models. The diagonal dotted line represents perfect calibration. (A) DeepSurv model; (B) Cox proportional hazards model; (C) random survival forest (RSF) model; and (D) eXtreme Gradient Boosting (XGBoost) model. The XGBoost model from the machine learning algorithm obtained a fairly satisfactory calibration, while the other models calibrated poorly. Supplementary Figure S3. The decision curve analysis at the 365th day of all models in the test dataset, evaluating clinical utility. (A) Decision curves, plotting net benefit against threshold probability. Net benefit is calculated as $(\text{True Positives}/n) - (\text{False Positives}/n) * (Pt/(1 - Pt))$, where Pt is the threshold probability and n is the total number of samples. (B) Net benefit of intervention avoidance curves, quantifying the net benefit of avoiding interventions based on model predictions. Supplementary Figure S4. The partial dependence plot (PDP) for the XGBoost model. It shows how the OS of the whole cohort changes if the value of one determinant is altered but all other factors are held constant. The y-axis represents the value of the survival function for each covariate. A wider area of the curve indicates that the greater the difference in levels of a factor, the greater the effect of that factor on OS. Supplementary Figure S5. The local explanations of all models, illustrating feature importance for individual predictions. (A and B) Local explanations provided by SurvLIME (Survival Local Interpretable Model-agnostic Explanations), showing the contribution of each feature to a specific prediction. (C) Local explanations provided by SurvSHAP(t) (Survival Shapley Additive Explanations (time-dependent)), showing the time-dependent feature attributions for individual predictions, quantifying how each feature influences the predicted survival function over time. Supplementary Table S1. Baseline characteristics of training data and test data.

Author Contributions: J.W., Q.K., S.T., S.Z., K.W. and G.F. contributed to the study conception and design. J.W. and K.W. contributed the material preparation, data collection, and analysis. J.W. and Q.K. wrote the first draft of the manuscript. G.F. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The studies involving humans were approved by the Human Institutional Review Board at the Beth Israel Deaconess Medical Center.

Informed Consent Statement: Informed consent was waived due to retrospective analysis of de-identified data from MIMIC-III (HIPAA-compliant public database).

Data Availability Statement: The datasets presented in this study can be obtained from the MIMIC database on the premise of completing its training (<https://physionet.org/content/mimiciv/3.1/>, accessed on 24 November 2024).

Acknowledgments: We thank the MIMIC-III team for data access, colleagues for technical discussions, and reviewers for enhancing manuscript quality.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Roger, V.L. Epidemiology of Heart Failure: A Contemporary Perspective. *Circ. Res.* **2021**, *128*, 1421–1434. [[CrossRef](#)] [[PubMed](#)]
2. Liu, Z.; Li, Z.; Li, X.; Yan, Y.; Liu, J.; Wang, J.; Guan, J.; Xin, A.; Zhang, F.; Ouyang, W.; et al. Global trends in heart failure from 1990 to 2019: An age-period-cohort analysis from the Global Burden of Disease study. *ESC Heart Fail.* **2024**, *11*, 3264–3278. [[CrossRef](#)]

3. Zannad, F.; Mebazaa, A.; Juillière, Y.; Cohen-Solal, A.; Guize, L.; Alla, F.; Rougé, P.; Blin, P.; Barlet, M.; Paolozzi, L.; et al. Clinical profile, contemporary management and one-year mortality in patients with severe acute heart failure syndromes: The EFICA study. *Eur. J. Heart Fail.* **2006**, *8*, 697–705. [\[CrossRef\]](#)
4. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2024**, *74*, 229–263. [\[CrossRef\]](#)
5. Levy, W.C.; Mozaffarian, D.; Linker, D.T.; Sutradhar, S.C.; Anker, S.D.; Cropp, A.B.; Anand, I.; Maggioni, A.; Burton, P.; Sullivan, M.D.; et al. The Seattle Heart Failure Model: Prediction of survival in heart failure. *Circulation* **2006**, *113*, 1424–1433. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Suzuki, S.; Yoshihisa, A.; Sato, Y.; Kanno, Y.; Watanabe, S.; Abe, S.; Sato, T.; Oikawa, M.; Kobayashi, A.; Yamaki, T.; et al. Clinical Significance of Get With the Guidelines-Heart Failure Risk Score in Patients With Chronic Heart Failure After Hospitalization. *J. Am. Heart Assoc.* **2018**, *7*, e008316. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Spinar, J.; Jarkovsky, J.; Spinarova, L.; Mebazaa, A.; Gayat, E.; Vitovec, J.; Linhart, A.; Widimsky, P.; Miklik, R.; Zeman, K.; et al. AHEAD score—Long-term risk classification in acute heart failure. *Int. J. Cardiol.* **2016**, *202*, 21–26. [\[CrossRef\]](#)
8. Jing, L.; Ulloa Cerna, A.E.; Good, C.W.; Sauers, N.M.; Schneider, G.; Hartzel, D.N.; Leader, J.B.; Kirchner, H.L.; Hu, Y.; Riviello, D.M.; et al. A Machine Learning Approach to Management of Heart Failure Populations. *JACC. Heart Fail.* **2020**, *8*, 578–587. [\[CrossRef\]](#)
9. Diprose, W.K.; Buist, N.; Hua, N.; Thurier, Q.; Shand, G.; Robinson, R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 592–600. [\[CrossRef\]](#)
10. Park, J.; Hwang, I.C.; Yoon, Y.E.; Park, J.B.; Park, J.H.; Cho, G.Y. Predicting Long-Term Mortality in Patients With Acute Heart Failure by Using Machine Learning. *J. Card. Fail.* **2022**, *28*, 1078–1087. [\[CrossRef\]](#)
11. Segar, M.W.; Jaeger, B.C.; Patel, K.V.; Nambi, V.; Ndumele, C.E.; Correa, A.; Butler, J.; Chandra, A.; Ayers, C.; Rao, S.; et al. Development and Validation of Machine Learning-Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multicohort Analysis. *Circulation* **2021**, *143*, 2370–2383. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Tian, P.; Liang, L.; Zhao, X.; Huang, B.; Feng, J.; Huang, L.; Huang, Y.; Zhai, M.; Zhou, Q.; Zhang, J.; et al. Machine Learning for Mortality Prediction in Patients With Heart Failure With Mildly Reduced Ejection Fraction. *J. Am. Heart Assoc.* **2023**, *12*, e029124. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Li, N.; Li, J.; Wang, K. Independent prognostic importance of the albumin-corrected anion gap in critically ill patients with congestive heart failure: A retrospective study from MIMIC-IV database. *BMC Cardiovasc. Disord.* **2024**, *24*, 735. [\[CrossRef\]](#)
14. Matsue, Y.; Kagiya, N.; Yamaguchi, T.; Kuroda, S.; Okumura, T.; Kida, K.; Mizuno, A.; Oishi, S.; Inuzuka, Y.; Akiyama, E.; et al. Clinical and Prognostic Values of ALBI Score in Patients With Acute Heart Failure. *Heart Lung Circ.* **2020**, *29*, 1328–1337. [\[CrossRef\]](#)
15. Li, N.; Li, J.; Wang, K. Association between red cell distribution width—Albumin ratio and all-cause mortality in intensive care unit patients with heart failure. *Front. Cardiovasc. Med.* **2025**, *12*, 1410339. [\[CrossRef\]](#)
16. Johnson, A.E.W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **2023**, *10*, 1, Erratum in *Sci. Data* **2023**, *10*, 219. [\[CrossRef\]](#)
17. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [\[CrossRef\]](#)
18. Riley, R.D.; Ensor, J.; Snell, K.I.E.; Harrell, F.E., Jr.; Martin, G.P.; Reitsma, J.B.; Moons, K.G.M.; Collins, G.; van Smeden, M. Calculating the sample size required for developing a clinical prediction model. *BMJ (Clin. Res. Ed.)* **2020**, *368*, m441. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Chen, Z.; Li, T.; Guo, S.; Zeng, D.; Wang, K. Machine learning-based in-hospital mortality risk prediction tool for intensive care unit patients with heart failure. *Front. Cardiovasc. Med.* **2023**, *10*, 1119699. [\[CrossRef\]](#)
20. Adler, E.D.; Voors, A.A.; Klein, L.; Macheret, F.; Braun, O.O.; Urey, M.A.; Zhu, W.; Sama, I.; Tadel, M.; Campagnari, C.; et al. Improving risk prediction in heart failure using machine learning. *Eur. J. Heart Fail.* **2020**, *22*, 139–147, Erratum in *Eur. J. Heart Fail.* **2020**, *22*, 2399. [\[CrossRef\]](#)
21. Shin, S.; Austin, P.C.; Ross, H.J.; Abdel-Qadir, H.; Freitas, C.; Tomlinson, G.; Chicco, D.; Mahendiran, M.; Lawler, P.R.; Billia, F.; et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Fail.* **2021**, *8*, 106–115. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Tong, R.; Zhu, Z.; Ling, J. Comparison of linear and non-linear machine learning models for time-dependent readmission or mortality prediction among hospitalized heart failure patients. *Heliyon* **2023**, *9*, e16068. [\[CrossRef\]](#) [\[PubMed\]](#)
23. 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. *J. Med. Internet Res.* **2022**, *24*, e38082. [\[CrossRef\]](#)
24. Hao, M.; Jiang, S.; Tang, J.; Li, X.; Wang, S.; Li, Y.; Wu, J.; Hu, Z.; Zhang, H. Ratio of Red Blood Cell Distribution Width to Albumin Level and Risk of Mortality. *JAMA Netw. Open* **2024**, *7*, e2413213. [\[CrossRef\]](#) [\[PubMed\]](#)

25. Wu, T.T.; Pan, Y.; Zheng, Y.Y.; Yang, Y.; Hou, X.G.; Deng, C.J.; Ma, Y.T.; Xie, X. Age-Bilirubin-International Normalized Ratio (INR)-Creatinine (ABIC) Score, a Potential Prognostic Model for Long-Term Mortality of CAD Patients After PCI. *J. Inflamm. Res.* **2023**, *16*, 333–341. [[CrossRef](#)]
26. Wang, J.; Wang, Y.; Feng, G.; Wang, K. Association between albumin corrected anion gap (ACAG) and all-cause mortality in intensive care unit heart failure patients treated with inotropes and vasopressors. *Signa Vitae* **2025**, *21*, 51–59. [[CrossRef](#)]
27. Wang, J.; Wang, K.; Feng, G.; Tian, X. Association Between the Albumin-Bilirubin (ALBI) Score and All-cause Mortality Risk in Intensive Care Unit Patients with Heart Failure. *Glob. Heart* **2024**, *19*, 97. [[CrossRef](#)]
28. Wang, J.; Li, N.; Mu, Y.; Wang, K.; Feng, G. Association between serum albumin creatinine ratio and all-cause mortality in intensive care unit patients with heart failure. *Front. Cardiovasc. Med.* **2024**, *11*, 1406294. [[CrossRef](#)]
29. Krzyżiński, M.; Spytek, M.; Baniecki, H.; Biecek, P. SurvSHAP(t): Time-dependent explanations of machine learning survival models. *Knowl. Based Syst.* **2023**, *262*, 110234. [[CrossRef](#)]
30. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.