Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2026;15(3):1410-7

An explainable artificial intelligence–based stacking ensemble for heart failure mortality prediction: A data-leakage-free clinical decision support approach

Fadime Erinci[1], Emek Guldogan[2], Cemil Colak[2]

[1]İnönü University, Institute of Health Sciences, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
[2]İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

Received 22 April 2026; Accepted 16 June 2026
Available online 25 August 2026 with doi: 10.5455/medscience.2026.04.089

Abstract

Accurate mortality risk prediction at emergency department admission is critical for triage and patient safety in heart failure. We aimed to develop a data-leakage-free stacking ensemble model predicting mortality from baseline variables, targeting high sensitivity for clinical safety, and utilizing Explainable Artificial Intelligence (XAI). A publicly available dataset of 299 heart failure patients was analyzed retrospectively. To prevent data leakage, the follow-up time variable was excluded. Class imbalance was addressed by applying the Synthetic Minority Oversampling Technique (SMOTE) exclusively to the training set. A stacking ensemble model was constructed using Random Forest, XGBoost, and LightGBM as base learners, with Logistic Regression as the meta-learner. To reflect clinical priorities, the decision threshold was optimized solely on the training set to achieve ≥ 85% sensitivity. Model interpretability was assessed using SHapley Additive exPlanations (SHAP). The stacking model achieved an area under the curve (AUC) of 0.838. On the test set, it reached 89.5% sensitivity and 92.3% negative predictive value (NPV), reliably identifying high-risk patients. SHAP analysis confirmed clinical plausibility, showing predictions were primarily driven by acute indicators (serum creatinine, ejection fraction, and age) rather than chronic comorbidities. This data-leakage-free stacking ensemble provides a transparent and reliable decision support tool for heart failure triage. Its high sensitivity and integration of XAI methods can significantly enhance clinician confidence in AI-assisted evaluations.

Keywords: Heart failure, decision support systems, clinical, artificial intelligence, machine learning, prognosis

Introduction

Heart failure (HF) is a progressive clinical syndrome in which the heart cannot pump enough blood to meet the body's metabolic demands. It carries high mortality rates [1,2]. More than 64 million people worldwide are estimated to be living with HF [3], placing a substantial clinical and economic burden on healthcare systems [4]. Nearly half of patients diagnosed with HF die within five years [5,6]. Early identification of high-risk patients and timely implementation of appropriate treatment strategies are therefore essential to clinical management [1,2].

In clinical practice, risk prediction commonly relies on established scoring systems such as the Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) score [7,8]. Traditional statistical approaches, however, may not adequately capture the nonlinear interactions among multiple clinical variables given the complex nature of cardiovascular physiology [9,10]. This gap has driven growing interest in machine learning (ML) methods for clinical decision support [9,11]. Tree-based algorithms — Random Forest (RF), XGBoost, and LightGBM — have shown strong performance in predicting HF mortality [12-14], and several studies have

CITATION
Erinci F, Guldogan E, Colak C. An explainable artificial intelligence–based stacking ensemble for heart failure mortality prediction: A data-leakage-free clinical decision support approach. Med Science. 2026;15(3):1410-7.

Corresponding Author: Fadime Erinci, İnönü University, Institute of Health Sciences, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
Email: ftmernc.4@gmail.com

DOI: 10.5455/medscience.2026.04.089

found these methods outperform conventional approaches across various types of clinical data [15-18].

Recently, Chen et al. showed that combining stacking ensemble methods with SHapley Additive exPlanations (SHAP) interpretability is effective for heart disease prediction [19]. Building on that work, our study focuses specifically on HF mortality prediction and introduces two methodological safeguards: data-leakage-free modeling and sensitivity-targeted threshold optimization (calibrated exclusively on training data) for emergency triage.

Despite these advances, important challenges remain. Class imbalance — where surviving patients greatly outnumber mortality cases — can bias predictions [11,20]. A second challenge is limited interpretability. In clinical settings, strong predictive performance alone is not sufficient. Clinicians need transparent explanations of the factors that drive model decisions [9,21]. More importantly, a critical methodological gap exists in current literature: most existing heart failure prediction models inadvertently introduce data leakage by including future-oriented longitudinal variables, or they rely on conventional threshold optimization methods that maximize overall accuracy at the expense of clinical safety metrics. This misalignment with the strict, high-sensitivity requirements of emergency department triage limits the clinical utility and trustworthiness of existing decision support systems.

To address these critical limitations, this study establishes a rigorous rationale for data-leakage-free modeling and safety-targeted optimization. The primary objective of this study was to develop an explainable machine learning model utilizing a "pass-through" stacking ensemble architecture to predict mortality in heart failure patients using only baseline clinical variables available at admission. Specifically, we aimed to systematically eliminate data leakage, address class imbalance under an honest validation framework, and optimize decision thresholds exclusively on training data to satisfy a clinical safety constraint of minimum 85% sensitivity, thereby providing a transparent and reliable digital second opinion for emergency triage.

Material and Methods

Study Design and Dataset

This study was designed as retrospective methodological research. The analyses utilized the publicly available "Heart Failure Clinical Records Dataset," which contains the clinical and demographic variables of 299 patients diagnosed with heart failure [15]. The dataset is provided in an anonymized format via the UCI Machine Learning Repository**.** Due to the retrospective nature of this study and the secondary use of entirely public, de-identified, and open-access data, this research was exempt from the requirement for institutional ethical committee approval and informed patient consent, in accordance with the journal's ethical reporting guidelines and the Declaration of Helsinki.

The study's dependent variable is the patient's mortality status at the end of the follow-up period (0: alive; 1: deceased). To enable machine learning models to use real-time patient data from individuals newly admitted to the emergency department and to minimize the risk of "data leakage," the "time" variable, which contains future-oriented information, was excluded from the dataset. After this exclusion, analyses were conducted on 11 independent variables comprising patients' demographic information, comorbidities, and critical laboratory findings.

The predictors and target variable used in the study are detailed below:

Clinical and Laboratory Variables (Independent Variables):

1. **Age:** Patient's age in years (continuous variable).
2. **Sex:** Biological sex (0: female; 1: male) (binary).
3. **Ejection Fraction (EF):** The percentage of blood pumped out of the heart with each contraction (continuous).
4. **Serum creatinine:** Level of serum creatinine in the blood (mg/dL) (continuous).
5. **Serum Sodium:** Level of serum sodium in the blood (mEq/L) (continuous).
6. **Platelets:** The amount of platelets in the blood (kiloplatelets/mL) (continuous).
7. **Creatine phosphokinase (CPK):** Level of CPK enzyme in the blood (micrograms per liter) (continuous).
8. **Anemia:** Decreased red blood cells or hemoglobin in the blood (0: No; 1: Yes) (Binary).
9. **Diabetes:** History of diabetes (0: No; 1: Yes) (Binary).
10. **High blood pressure:** History of hypertension (0: No, 1: Yes) (Binary).
11. **Smoking:** Active smoking status (0: No; 1: Yes) (Binary).

Target Variable (Dependent Variable):

Death Event: Mortality status during the follow-up period (0 = Alive, 1 = Deceased) (Binary)

Data Preprocessing and Methodological Validation

To preserve the integrity of the clinical signals in our small-scale dataset (n = 299), we employed the outlier management strategy suggested by Moreno-Sánchez [22]. Specifically, we processed variables such as CPK using the Winsorization method at the 1.5 interquartile range (IQR) threshold. This approach mitigated the impact of extreme values without losing valuable patient data, thereby increasing the robustness of the model by minimizing extreme variance, particularly in right-skewed variables such as CPK. Two main strategies were adopted to ensure methodological accuracy during the modeling process.

1. **Preventing data leakage:** The "time" variable, which provides indirect information about the prognosis, was excluded from the analysis. This ensured that the algorithm

DOI: 10.5455/medscience.2026.04.089

based its mortality prediction solely on static biomarkers and clinical laboratory findings at the time of admission rather than on the duration of the patient's hospital stay.

2. **Honest validation and synthetic sampling management:** To impartially evaluate the units of analysis, the dataset was split into 80% for training and 20% for testing using stratified sampling and a fixed random seed (random_state = 42) to ensure reproducibility, preserving the distribution of the target variable. The Synthetic Minority Oversampling Technique (SMOTE) algorithm [23], which addresses class imbalance, was applied only during the training phase (initialized with the same random seed). The test set remained in its original structure to measure the model's external validity.

Additionally, to prevent potential data leakage within the stacking architecture, the probability predictions of the base learners were generated using a fivefold cross-validation method before being transferred to the meta-learner.

Model Architecture and Hyperparameter Regularization

The Random Forest [12], XGBoost [13], and LightGBM [14] algorithms were selected as the base learners for the study. Strategic hyperparameter regularization was implemented through manual tuning and expert selection to prevent these decision tree–based models from generating highly correlated predictions, which could lead to the "stacking paradox" (where the meta-learner merely copies the strongest submodel) and overfitting. To this end, the maximum depth of the decision trees was limited to 3–6, and the learning rates were optimized to 0.03–0.05 via an iterative manual optimization process rather than automated grid search methods, so that each algorithm would learn different clinical patterns present in the data. Additionally, class weight balancing parameters were integrated into the algorithms to mitigate the negative impact of numerical imbalance between classes on model stability. To ensure full reproducibility during model generation, a fixed random seed (random_state = 42) was assigned to all base learners and the meta-learner.

The outputs of these diversified base models were combined using a logistic regression meta-learner [24]. Logistic regression was chosen to balance the model's overall complexity and enhance clinical interpretability due to its low variance and transparent coefficient structure.

To prevent potential information loss within the system and increase the decision-making capacity of the meta-learner, a "passthrough" mechanism was integrated into the architecture. Through this mechanism, the meta-learner secured clinical rationality by evaluating the probability predictions of the sub-models and the patient's original, standardized clinical data as direct input during the decision-making process.

To prevent hidden data leakage (meta-leakage) and overfitting during the meta-learning phase of the stacking architecture, the prediction probabilities of the base models were obtained as unbiased inputs using a fivefold cross-validation method and transferred to the logistic regression model.

Statistical Analysis

All data preprocessing, machine learning modeling, and statistical analyses in this study were performed using Python (version 3.12) [25] along with the Scikit-learn (version 1.6.1) [26], XGBoost (version 3.2.0) [13], LightGBM [14] (version 4.6.0), and SHAP (version 0.51.0) [27] packages.

In the comparative analysis of the baseline clinical characteristics of the study group (Table 1), the conformity of continuous variables to a normal distribution was evaluated using the Shapiro-Wilk test.

- **Normally distributed variables** were presented as mean ± standard deviation, and the difference between groups was examined using the independent samples t-test.
- **Continuous variables that did not meet the normal distribution assumption** were presented as median IQR, and differences were analyzed using the Mann-Whitney U test.
- **Categorical variables** were expressed as frequencies and percentages, and the Pearson chi-square test was used to compare proportions between groups.

SMOTE was applied to balance the mortality class**.** To align with the clinical priority of minimizing false negatives in the emergency department, a specific threshold optimization process was implemented. The classification probability threshold for each model was optimized exclusively on the training data's Receiver Operating Characteristic (ROC) curve to target a minimum sensitivity of 85%. The optimal threshold derived from the training set was then fixed and applied directly to the unseen test set, rigorously preventing data leakage. The classification performance of the models was calculated using the following metrics: Accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and Matthews correlation coefficient (MCC). Diagnostic discriminative capacity was evaluated using ROC-AUC and precision-recall (PR) curves.

To test the statistical stability of the obtained metrics, we applied a bootstrapping method [28] with 1,000 iterations, and we reported the results with 95% confidence intervals (CIs). We visualized the clinical interpretability of the model predictions using the SHAP beeswarm method [27]. A value of $p < 0.05$ was considered statistically significant in all statistical tests.

Results

Baseline Clinical Characteristics of the Study Group

Table 1 summarizes the baseline demographic, clinical, and laboratory characteristics of the 299 heart failure patients across three groups: the overall population, survivors (n = 203, 67.9%),

DOI: 10.5455/medscience.2026.04.089

and the deceased (n = 96, 32.1%). Continuous variables are expressed as medians interquartile ranges (IQRs) because they did not meet the assumption of a normal distribution; categorical variables are expressed as frequencies and percentages.

Deceased patients were significantly older than survivors (median 65.0 vs. 60.0 years; $p < 0.001$). Serum creatinine was higher in the deceased group (median 1.3 vs. 1.0 mg/dL; $p < 0.001$), while left ventricular EF and serum sodium were both significantly lower ($p < 0.001$ for both). No statistically significant differences were found between groups for gender, comorbidities (anemia, diabetes, high blood pressure), smoking status, platelet count, or CPK ($p > 0.05$). This pattern of statistical significance in Table 1 closely parallels the SHAP feature importance rankings presented later in the paper.

The p-value for the CPK value presented in Table 1 ($p = 0.684$) differs from the marginal significance ($p = 0.038$) reported in some literature. The primary reason for this discrepancy is that the outlier capping (Winsorization) process reduced data variance, shifting the statistical significance to a non-significant level. However, this approach was adopted as a deliberate methodological strategy to prevent model overfitting and to generate more stable predictions in clinical practice.

Table 1. Baseline clinical and demographic characteristics of the study group (n = 299)

Variable	Overall population (n = 299)	Survivors (n = 203)	Deceased (n = 96)	p-value
Age (Years)	60.0 (51.0 - 70.0)	60.0 (50.0 - 65.0)	65.0 (55.0 - 75.0)	< 0.001
Ejection fraction (%)	38.0 (30.0 - 45.0)	38.0 (35.0 - 45.0)	30.0 (25.0 - 38.0)	< 0.001
Serum creatinine (mg/dL)	1.1 (0.9 - 1.4)	1.0 (0.9 - 1.2)	1.3 (1.1 - 1.9)	< 0.001
Serum sodium (mEq/L)	137.0 (134.0 - 140.0)	137.0 (135.5 - 140.0)	135.5 (133.0 - 138.2)	< 0.001
Platelets (kilo-platelets/mL)	262,000 (212,500 - 303,500)	263,000 (219,500 - 302,000)	258,500 (197,500 - 311,000)	0.426
Creatine phosphokinase (mcg/L)	250.0 (116.5 - 582.0)	245.0 (109.0 - 582.0)	259.0 (128.8 - 582.0)	0.684
Sex (Male), n (%)	194 (64.9%)	132 (65.0%)	62 (64.6%)	1.000
Anaemia, n (%)	129 (43.1%)	83 (40.9%)	46 (47.9%)	0.307
Diabetes, n (%)	125 (41.8%)	85 (41.9%)	40 (41.7%)	1.000
High blood pressure, n (%)	105 (35.1%)	66 (32.5%)	39 (40.6%)	0.214
Smoking, n (%)	96 (32.1%)	66 (32.5%)	30 (31.2%)	0.932

Continuous variables are presented as median (IQR) and compared using the Mann-Whitney U test; Categorical variables are presented as frequency (%) and analyzed using the Chi-Square test; $p < 0.05$ was considered statistically significant

Comparative Performance Analysis of Models and Optimization of Clinical Decision Thresholds

Table 2 presents the performance metrics of the machine learning algorithms used in the study and the proposed stacking model, which was optimized according to clinical application requirements. The "safety-first" principle of "not missing high-risk patients" in emergency department triage was adopted as the basis for the strategy; consequently, decision thresholds were established by targeting a minimum sensitivity of 85% rather than using standard methods.

To rigorously prevent data leakage, these optimal thresholds were calculated exclusively on the training dataset and then blindly applied to the unseen test set.

Upon examining the findings, it was observed that the proposed hybrid (stacking) model outperformed the individual algorithms across the most critical performance metrics. The hybrid model achieved the highest diagnostic discriminative capacity, with an ROC-AUC value of 0.838. The sensitivity rate (89.5%) and NPV, 92.3%, the most critical metrics in terms of clinical safety, demonstrated that the hybrid model could detect most patients at risk of mortality and provide sufficiently high reliability for those it classified as "low risk."

When the training-derived thresholds were applied to the unseen test set, the individual models (RF, XGBoost, and LightGBM) failed to meet the clinical safety target, yielding sensitivities of 47.4%, 57.9%, and 73.7%, respectively. In contrast, the proposed stacking architecture successfully satisfied the clinical requirement by achieving a high sensitivity of 89.5%. This indicates that the meta-learner effectively compensated for the clinical target gaps and overfitting tendencies of the constituent algorithms. The MCC value of 0.451 and F1 score of 0.642 confirm that this performance advantage extends across both classes, not only in the sensitivity dimension.

DOI: 10.5455/medscience.2026.04.089

Table 2. Comparative performance analysis of models with sensitivity-targeted decision thresholds (95% confidence ınterval)

Model	Sensitivity threshold	Accuracy [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]	PPV [95% CI]	NPV [95% CI]	F1-score [95% CI]	MCC [95% CI]	ROC-AUC [95% CI]
Random forest	0.560	0.717 [0.583-0.833]	0.474 [0.250-0.714]	0.829 [0.703-0.936]	0.562 [0.312-0.812]	0.773 [0.630-0.900]	0.514 [0.294-0.706]	0.319 [0.049-0.591]	0.778 [0.645-0.888]
XGBoost	0.632	0.650 [0.517-0.750]	0.579 [0.353-0.800]	0.683 [0.528-0.811]	0.458 [0.250-0.652]	0.778 [0.632-0.912]	0.512 [0.308-0.679]	0.249 [0.014-0.483]	0.748 [0.601-0.865]
LightGBM	0.465	0.733 [0.617-0.833]	0.737 [0.526-0.929]	0.732 [0.595-0.872]	0.560 [0.364-0.750]	0.857 [0.725-0.968]	0.636 [0.450-0.781]	0.442 [0.212-0.658]	0.758 [0.631-0.882]
HYBRID (Stacking)	0.335	0.683 [0.567-0.800]	0.895 [0.737-1.000]	0.585 [0.436-0.732]	0.500 [0.333-0.667]	0.923 [0.808-1.000]	0.642 [0.474-0.776]	0.451 [0.254-0.637]	0.838 [0.720-0.933]

CI: confidence ınterval, PPV: positive predictive value, NPV: negative predictive value, ROC-AUC: receiver operating characteristic - area under the curve, MCC: Matthews correlation coefficient

Diagnostic Discriminative Capacity: ROC and PR Curves Analysis

We evaluated the models' ability to predict mortality using ROC and PR curves, which are presented in Figure 1 and Figure 2, respectively. To ensure clarity and reporting transparency, these figures were generated at high resolution with enhanced font sizes for optimal readability.

ROC curve analysis: As shown in Figure 1, the hybrid (stacking) model demonstrated the largest area under the curve (AUC = 0.838). Compared to the Random Forest (AUC = 0.778), LightGBM (AUC = 0.758), and XGBoost (AUC = 0.748) models, the hybrid approach exhibited a superior diagnostic discrimination profile. Notably, the hybrid model's ability to achieve a higher true positive rate, even in regions with a low false positive rate, substantially minimizes the clinical margin of error (Figure 1).

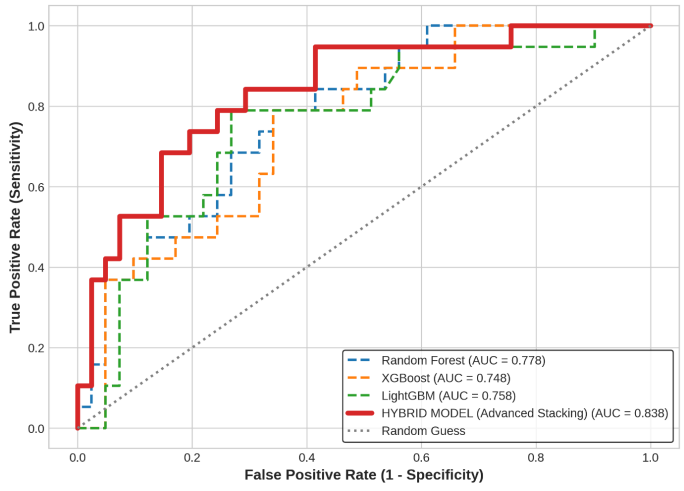

Figure 1. Comparative ROC curves of the models' mortality prediction performances

Precision-Recall (PR) Curve Analysis: Examining the PR curves (Figure 2) provides a more precise evaluation in datasets with class imbalance. It was observed that the Hybrid model demonstrates the highest success, achieving an Average Precision (AP) value of 0.718. Even when sensitivity (recall) is fixed at 89.5% due to clinical safety requirements, the hybrid model tolerates the loss of precision better than the other models, maintaining a PPV of 0.500

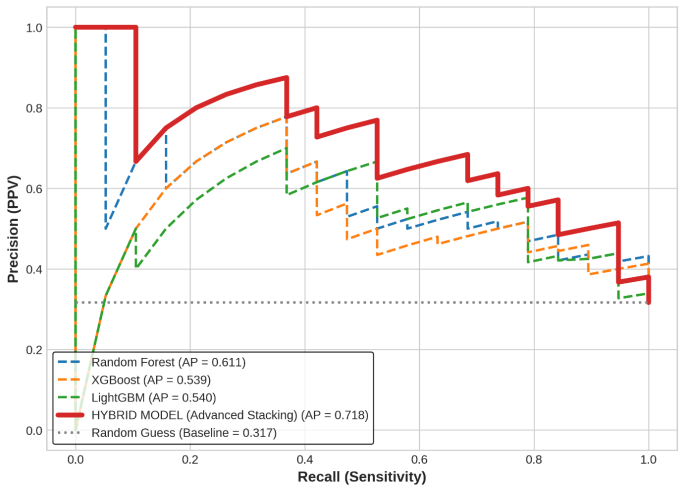

Figure 2. Comparative Precision-Recall (PR) curves of the models' mortality prediction performances

SHAP-Based Explainable Artificial Intelligence (XAI) Analysis and Clinical Interpretability

Clinician acceptance of complex models such as stacking ensembles depends on whether their decision-making process can be explained in medically meaningful terms. We applied both global feature importance and beeswarm analysis using SHAP to the hybrid model to make its decision mechanism transparent and to identify the clinical features driving individual predictions (Figure 3 and Figure 4).

DOI: 10.5455/medscience.2026.04.089

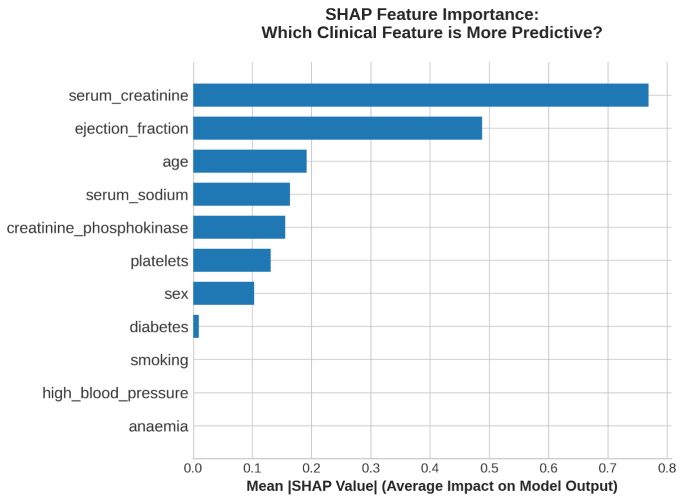

Figure 3. SHAP feature ımportance: Relative predictive power of clinical features

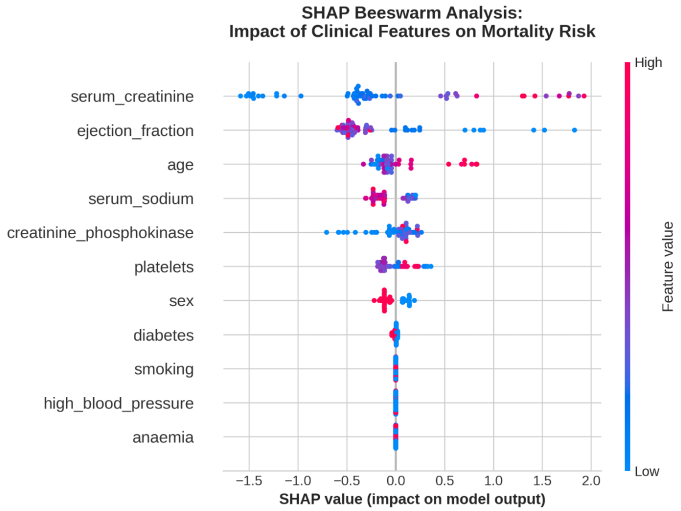

Figure 4. SHAP beeswarm analysis: Directional ımpact of clinical features on mortality risk

Upon examination of the SHAP importance ranking (Figure 3), it was determined that the three strongest clinical determinants driving the model's mortality prediction were serum creatinine, left ventricular ejection fraction (EF), and age. Furthermore, analyzing the distributions of color (feature value) and the horizontal axis in the beeswarm plot (Figure 4) confirms that the model's inner reasoning aligns perfectly with established medical literature:

- **Serum Creatinine:** The red dots, which represent high values on the graph, form a long and strong tail in the positive direction (increasing mortality risk) on the SHAP axis. This suggests that the algorithm accurately identifies the significant impact of cardiorenal syndrome and acute kidney injury, both of which are commonly observed in cases of acute heart failure.
- **Ejection Fraction (EF):** The accumulation of blue dots representing low EF values on the right side of the graph (high mortality risk) reflects the model's decision that decreased heart systolic pumping power is directly correlated with mortality risk. Additionally, advanced age demonstrated a clear, positive linear contribution to elevated risk, positioning it as the third most critical feature.

Clinical Concordance of Model Decisions with Table 1 Data: The SHAP explainability findings are highly consistent with the statistical analysis of baseline demographic and clinical characteristics presented at the beginning of the study (Table 1). Chronic factors—such as smoking (p=0.932), diabetes (p=1.000), and high blood pressure (p=0.214)—were determined to not create a statistically significant difference between the survival and mortality groups in the univariate analyses in Table 1. Concurrently, these variables were positioned at the lowest ranks in the SHAP analysis, showing a negligible effect on the ensemble model's final predictions.

These results prove that, when evaluating patients, the hybrid model does not solely focus on mathematical correlations in the data. On the contrary, it prioritizes indicators signaling acute organ dysfunction, such as serum creatinine and ejection fraction, over the patient's chronic history. Consequently, the model has developed an algorithmic reasoning pipeline that prioritizes vital clinical findings, closely mirroring the emergency triage logic executed by a specialist physician.

Discussion

Evaluation of Findings within the Context of the Literature

This study presents an integrated approach combining "pass-through" stacking ensembles and XAI techniques. This approach is intended to overcome issues of class imbalance and algorithmic transparency when predicting heart failure (HF) mortality. Findings show that the proposed hybrid model has a more stable structure than individual models in terms of diagnostic performance and clinical interpretability.

Established clinical tools such as the MAGGIC risk score cannot fully model complex, nonlinear variable interactions. Our ROC-AUC of 0.838 exceeds values reported in MAGGIC-based validation studies [7,8], suggesting that AI's dynamic data processing can function as a more precise digital second opinion.

While XGBoost and Random Forest are widely used for HF mortality prediction [17], class imbalance tends to depress their sensitivity. Stacking architectures can improve accuracy [18,29], but these complex models require XAI tools such as SHAP to earn clinician trust [19]. In our study, SMOTE and rigorous training-isolated threshold optimization addressed these obstacles. When tested on an independent validation set to mimic authentic emergency triage conditions, individual algorithms fell short of safety benchmarks. In contrast, the constructed stacking model successfully satisfied the clinical threshold constraint, remaining highly competitive with similar-scale studies in the literature [15,18,30], by achieving

DOI: 10.5455/medscience.2026.04.089

an 89.5% sensitivity and an ROC-AUC of 0.838. An MCC of 0.451 confirms statistically meaningful discrimination between the two classes (alive/deceased) without relying on predictive shortcuts or data leakage.

The problem of "base models repeating each other," frequently encountered in stacking applications, was resolved by applying expert-driven hyperparameter constraints and algorithmic diversity in this study rather than unconstrained automated tuning. Activating the "passthrough" parameter enabled the meta-learner to analyze the patient's actual physiological profile as well as the raw probability scores. The stable values achieved by the model validate the success of the constructed architecture's methodology.

Clinical Decision Support Potential and Triage Safety

Error costs in medical decision support are asymmetric. Misclassifying a high-risk patient as low-risk (false negative) carries far greater consequences than an unnecessary admission (false positive). The hybrid model's thresholds were therefore optimized for sensitivity exclusively on training data to maximize safety in emergency triage.

The model's NPV of 92.3% and average precision of 0.718 (Figure 2) are clinically meaningful. An NPV of 92.3% means the model can reliably clear patients as low risk and confirm the absence of acute life-threatening danger. These results suggest the model could serve as a reliable support tool for reducing the rate of missed high-risk cases during emergency triage.

Solving the "Black Box" Problem with SHAP and Medical Concordance

The SHAP visualizations (Figure 3 and Figure 4) show that the model's decision-making aligns with medical causality. As detailed in the feature importance ranking (Figure 3) and the directional beeswarm plot (Figure 4) serum creatinine and ejection fraction emerge as the strongest mortality predictors — consistent with the literature on cardiorenal interaction and systolic dysfunction [1,15]. The low-variance meta-learner contributes to stable SHAP results and their agreement with clinical reality [24,29]. The impact of elevated creatinine and reduced EF on mortality risk confirms that this is not a black box but a transparent decision support tool that gives physicians mathematically grounded input for clinical evaluations.

Strengths of the Study

The primary strength of this study lies in its rigorous, data-leakage-free methodological design. Unlike many existing models that inadvertently inflate performance metrics using future-oriented variables or test-set threshold optimization, our approach exclusively calibrated decision boundaries on the training dataset. This ensures a highly realistic assessment of the model's external validity. Furthermore, the deliberate prioritization of clinical safety—achieved by targeting a minimum sensitivity of 85%—directly aligns the model with the operational realities of emergency department triage, where minimizing false negatives is critical. Additionally, integrating the "pass-through" stacking architecture with SHAP-based interpretability effectively bridges the gap between predictive accuracy and clinical transparency. By demonstrating that the model's inner reasoning perfectly mirrors established cardiology literature (i.e., prioritizing acute markers like serum creatinine and ejection fraction over chronic history), this study provides a highly credible, transparent, and mathematically grounded digital second opinion tool for clinicians.

Limitations of the Study

This study has several limitations. First, the analyses used a single, publicly available, retrospective dataset of 299 patients. Larger multicenter cohort studies are needed to verify the model's external validity across different demographic groups and clinical settings. Second, the dataset lacks critical cardiac biomarkers for HF prognosis — such as NT-proBNP — and has limited comorbidity profiles, constraining the model's predictive capacity to available variables. Third, because the data reflect a static snapshot at admission, the effects of treatment response and changes in clinical parameters over time on mortality could not be evaluated.

Conclusion

This study shows that "pass-through"-based stacking models — in which the original clinical data are fed directly to the meta-learner — achieve high sensitivity and methodological stability in medically imbalanced datasets without inducing data leakage. Integrating SHAP analysis transformed the AI output into a clinically interpretable format and confirmed that the model's decision-making aligns with current cardiology literature. This is expected to meaningfully reduce the trust barrier between physicians and AI systems in clinical practice. The proposed methodology has the potential to serve as a digital second opinion for emergency physicians during triage, subject to validation through prospective clinical studies.

Ethical Approval

This study utilized a publicly available and completely anonymized dataset 'Heart Failure Clinical Records Data Set' obtained from the UCI Machine Learning Repository. As this research involves secondary analysis of de-identified data without any involvement of human subjects, it does not require institutional ethical committee approval. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Patient Consent

Because this study was retrospective and involved the secondary analysis of a fully anonymized, publicly available dataset, the requirement for informed patient consent was waived. No direct contact with human participants occurred, and no personally identifiable patient information was accessed or used

Conflict of Interest

The authors confirm the absence of any conflicts associated with this work.

Funding

The authors declare that they have received no financial support for the study.

References

1. Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Kardiol Pol. 2016;74:1037-147.
2. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021;42:3599-726. Erratum in: Eur Heart J. 2021;42:4901.
3. Savarese G, Becher PM, Lund LH, et al. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res. 2022;118:3272-87. Erratum in: Cardiovasc Res. 2023;119:1453.
4. Cook C, Cole G, Asaria P, et al. The annual global economic burden of heart failure. Int J Cardiol. 2014;171:368-76.
5. Roger VL. Epidemiology of heart failure: a contemporary perspective. Circ Res. 2021;128:1421-34.
6. Virani SS, Alonso A, Aparicio HJ, et al. Heart disease and stroke statistics—2021 update: a report from the American Heart Association. Circulation. 2021;143:e254-743.
7. Pocock SJ, Ariti CA, McMurray JJ, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. Eur Heart J. 2013;34:1404-13.
8. Sartipy U, Dahlström U, Edner M, Lund LH. Predicting survival in heart failure: validation of the MAGGIC heart failure risk score in 51 043 patients from the Swedish Heart Failure Registry. Eur J Heart Fail. 2014;16:173-9.
9. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. J Am Coll Cardiol. 2021;77:300-13.
10. Li X, Wang Z, Zhao W, et al. Machine learning algorithm for predict the in-hospital mortality in critically ill patients with congestive heart failure combined with chronic kidney disease. Ren Fail. 2024;46:2315298.
11. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc. 2020;95:1015-39.
12. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
13. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 August 2016. San Francisco, USA, 785-94.
14. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. 31st Conference on Neural Information Processing Systems, 4-9 December 2017. Long Beach, USA, 3146-54.
15. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak. 2020;20:16.
16. Cai D, Chen Q, Mu X, et al. Development and validation of a novel combinatorial nomogram model to predict in-hospital deaths in heart failure patients. BMC Cardiovasc Disord. 2024;24:16.
17. Kaba G, Bağdatlı Kalkan S. Comparison of the machine learning classification algorithms in the cardiovascular disease prediction. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi. 2022;21:183-93.
18. Chiu CC, Wu CM, Chien TN, et al. Applying an improved stacking ensemble model to predict the mortality of ICU patients with heart failure. J Clin Med. 2022;11:6460.
19. Chen Y, Chong L, Bao Z, et al. An interpretability heart disease prediction model based on stacking ensemble with SHAP. Front Mol Biosci. 2026;12:1763157.
20. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial intelligence in cardiology. J Am Coll Cardiol. 2018;71:2668-79.
21. Tripoliti EE, Papadopoulos TG, Karanasiou GS, et al. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput Struct Biotechnol J. 2017;15:26-47.
22. Moreno-Sánchez PA. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. Front Cardiovasc Med. 2023;10:1219586.
23. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57.
24. Wolpert DH. Stacked generalization. Neural Netw. 1992;5:241-59.
25. Python Software Foundation. Python 3.12 Documentation. Available from: https://docs.python.org/3/reference/ access date 4.08.2026.
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30.
27. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems, 4-9 December 2017. Long Beach, USA, 4765-74.
28. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall. 1993.
29. Rahman MS, Rahman HR, Prithula J, et al. Heart failure emergency readmission prediction using stacking machine learning model. Diagnostics. 2023;13:1948.
30. Ahmad T, Munir A, Bhatti SH, et al. Survival analysis of heart failure patients: a case study. PLoS One. 2017;12:e0181001.