Big Data and Information Analytics

BDIA, 8: 43–64.
DOI: 10.3934/bdia.2024003
Received: 08 June 2024
Revised: 21 July 2024
Accepted: 30 July 2024
Published: 08 August 2024

https://www.aimspress.com/journal/BDIA

Research article

Motality prediction of ICU rheumatic heart disease with imbalanced data based on machine learning

Yiwen Tao[1], Zhenqiang Zhang[2], Bengbeng Wang[3] and Jingli Ren[1,*]

[1] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China
[2] School of Medicine, Henan University of Chinese Medicine, Zhengzhou 450003, China
[3] BYD Company Ltd, Shenzhen 518118, China

* **Correspondence:** Email: renjl@zzu.edu.cn.

Abstract: Linked to poverty, rheumatic heart disease (RHD) disproportionately burdens the developing world, receiving less attention than other infectious diseases. Resampling and cost-sensitive learning techniques are applied to predict the mortality risk of imbalanced RHD datasets. A total of 57 models were constructed, and was comprised of 50 resampled machine learning (ML) models and 7 cost-sensitive learning models. The results from the Friedman and Nemenyi tests highlight the superior performance of the cost-sensitive support vector classification model, with an AUC of 0.888, sensitivity of 0.800, G-means of 0.806, and a Brier score of 0.061. The global and local interpretability are advanced through two post-hoc interpretable ML methods, facilitating the prioritization of key features associated with mortality risk, the determination of thresholds for features, and a comprehension of how variations in these features influence patient mortality rates. These findings may prove to be clinically valuable, assisting clinicians in tailoring precise management that is essential to maximize the survival of RHD patients.

Keywords: rheumatic heart disease; explainable machine learning; prediction mortality; resampling and cost-sensitive learning; imbalanced data

1. Introduction

Rheumatic heart disease (RHD) is a cardiac condition that stems from rheumatic inflammation, leading to damage in the heart valves. The primary instigator of heart valve disease is rheumatic fever,

which is initiated by an infection with group A beta-hemolytic streptococcus [1,2]. This infection primarily inflicts harm on the heart valves, which results in stenosis and an insufficient closure, and ultimately progresses to heart failure [3]. The recent Global Burden of Diseases report on RHD serves as a timely reminder of the notable global heterogeneity in RHD burden [4,5]. In developed nations, the incidence of RHD has significantly dropped due to better living conditions, widespread healthcare education, and the widespread use of penicillin. However, in certain regions and among vulnerable groups, especially in low and middle-income countries such as Africa, the Western Pacific, and India, RHD continues to be prevalent at high levels. This is mainly because of limited healthcare resources and difficulties in early diagnoses and treatments [5].

Accurately predicting the in-hospital mortality risk of patients in intensive care units (ICU) is crucial to optimize treatment plans, to make informed clinical decisions, and to establish harmonious doctor-patient relationships [6]. Currently, statistical models that utilize serum related markers such as IL-1β, IL-8, IL-6, tumor necrosis factor α, and anti-streptolysin for the prognostic assessment of RHD are in clinical use [7,8]. However, owing to the limitations of these statistical methods in addressing intricate relationships between clinical and biological factors, the prognostic results frequently suffer from a lack of specificity or sensitivity [8–10].

In recent years, with significant advancements in computer performance, the integration of medical engineering has become increasingly prominent, and machine learning (ML) algorithms have progressively entered the field of medicine [7]. Ngiam et al. demonstrated that Ml algorithms outperformed traditional methods (e.g., descriptive statistics, inferential statistics, non-parametric statistics), especially in evaluating critically ill patients with severe conditions and extensive, complex clinical data [11]. ML approaches actively seek to capture various rich and interesting features by considering multidimensional nonlinear patterns among variables, regardless of their complexity. Consequently, ML algorithms have been widely applied to prognostic assessments across diverse patient populations [12–14]. In the domain of RHD, prior research has primarily focused on the utilization of ML algorithms for diagnoses [15,16] and classification [17]. Regarding mortality prediction for RHD patients in the ICU, to our knowledge, there are only a few existing studies [10]. This specific study employed XGBoost and a logistic regression, and exhibited a commendable performance in the development of prognostic models.

The imbalance between major and minor classes often results in frequent errors in prediction and classification. In medical datasets, the minor class typically refers to events with lower occurrence rates, such as disease recurrence, progression, or mortality [18]. Errors in predicting the minor class have a more substantial impact on the clinical outcomes compared to errors in predicting the major class. For instance, misclassifying disease progression as normal or low-risk may inadvertently categorize patients as normal or low-risk, respectively, leading to potential adverse consequences [19–21]. The age-standardized death rate for RHD was approximately 3.9 per 100,000 in 2019 [22], implying the potential occurrence of imbalanced data in the RHD dataset. However, a pioneering effort to construct predictive mortality models seemed to have overlooked this aspect in their work [10]. Therefore, our paper focuses on examining the issue of data distribution uniformity to optimize the predictive performance of mortality risk in RHD patients.

Moreover, despite the favorable performance of ML algorithms in previous studies, the inherent "black-box" nature of ML algorithms makes it challenging to elucidate which patient features are accountable for a given prediction. The lack of interpretability has been a significant impediment to the adoption of ML models in the medical domain [23]. To interpret the results of ML models, our

paper integrates ML algorithms with post-hoc interpretable ML techniques such as Shapley additive explanations (SHAP; [24]) and local interpretable model-agnostic explanations (LIME; [25]). This integration aims to provide a deeper understanding of the complex relationships between features and predictions.

To address the gaps that arise from data imbalance and the lack of interpretability when predicting the death risk for RHD patients, this paper aims to complete the following:

(1) Develop RHD mortality risk predictive models customized for imbalanced data distribution;

(2) Enhance the interpretability of the constructed ML model, revealing the key features that influence the mortality risk of RHD patients and elucidating how variations in these features impact changes in mortality risk.

In addition to optimizing the predictive performance of mortality risk in critically ill RHD patients, this study also offers intuitive explanations. These explanations assist clinicians in understanding the specific prediction process of the developed model, facilitate the early identification of high-risk individuals for in-hospital mortality, and increase opportunities for early intervention. Consequently, this facilitates the optimization of treatment plans and the formulation of clinical decisions that maximize the benefits for patients.

2. Materials and methods

2.1. Data

In this paper, we utilized the freely accessible critical care database known as the Medical Information Mart for Intensive Care (MIMIC-IV) database version 4.1, with the necessary permission granted (certificate number: 48369375). The MIMIC-IV database contains comprehensive clinical data pertaining to patients admitted to the Beth Israel Deaconess Medical Center between 2008 and 2019. It encompasses a vast dataset, including over 200,000 emergency department admissions and more than 70,000 ICU stays. The clinical data within MIMIC-IV encompasses a wide range of information, including demographic characteristics, vital signs, results from imaging examinations, laboratory tests, a data dictionary, and documents containing codes from the International Classification of Diseases, Ninth and Tenth Revisions (ICD-9 and ICD-10, respectively). Additionally, it contains records of hourly physiologic data obtained from bedside monitors, which have undergone validation by ICU nurses. Crucially, it's important to note that all health information sourced from the MIMIC-IV database is anonymized, thus eliminating the need for informed consent from the patients involved. This database has received approval from the Institutional Review Boards of the Massachusetts Institute of Technology.

To filter missing data, we employed the missingno module in the Python software, version 3.9.12. In Figure 1(a), each column represents a clinical variable, and the white line signifies missing data. The denser the white lines within each column, the greater the number of missing values for that variable. Patients that met specific criteria were chosen for this study from the database, including the following: (1) their initial ICU admission occurring during their first hospitalization, (2) an ICU length of stay exceeding 24 hours, and (3) an age of over 18 years. After removing 452 cases with a missing rate that exceeded 80% for laboratory indicators and 507 cases with a total indicator missing rate that exceeded 20%, there remained 1266 study samples. This cohort comprised 1150 in-hospital survivors and 116 in-hospital deaths. The identification of patients with RHD was carried out using ICD-9 and

ICD-10 codes. Each case encompassed 163 data items, and covered demographic information, vital signs recorded within 24 hours of ICU admission, and laboratory results. The data extraction process is depicted in Figure 1(b). In Figure 1(c), we present the distribution of inpatient statuses in the research cohort, with 91% of the population being classified as survivors and 9% as deceased. Regarding the gender distribution, females account for 49%, while males make up 51%. In terms of the racial distribution, the majority are White, comprising 67.77%, followed by African Americans at 8.69%, other races at 5.53%, Hispanic/Latinx at 3.55%, Asians at 3.24%, an unknown race at 11.06%, and American Indian/Alaska Native at 0.16%. Furthermore, we conducted an analysis of the age distribution across different statuses. Missing values were imputed using the K-nearest neighbor (KNN) interpolation method [16].

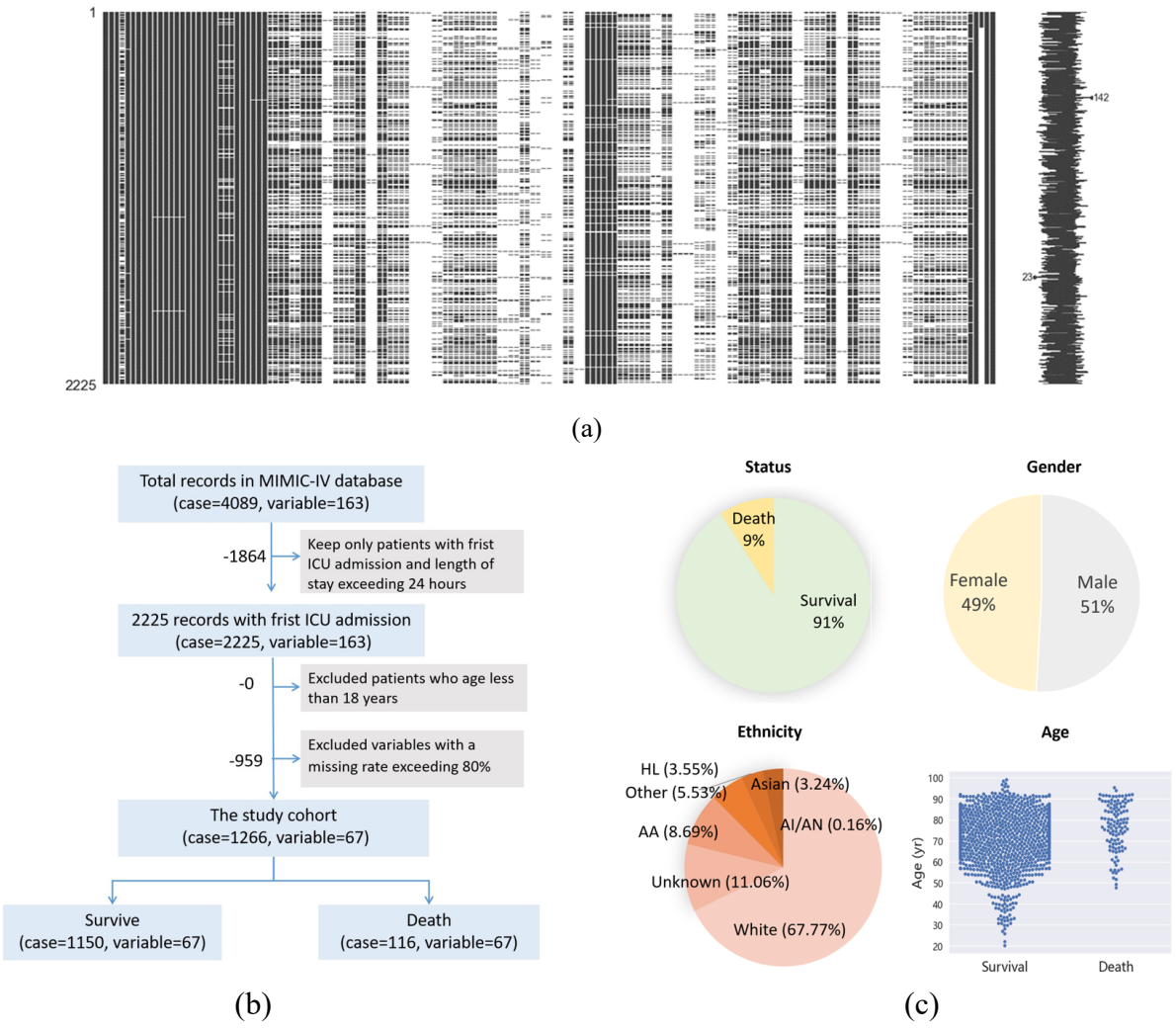

Figure 1. (a) Visualization for variable missingness: the white lines represent missing data. (b) The detailed process of data extraction. (c) The status, gender, ethnicity and age distribution of the study cohort.

2.2. Statistical analysis

The Friedman test [27] is a non-parametric statistical test used to analyze data in which multiple related groups or conditions are compared. The test works by ranking the data within each group, calculating the average rank for each group, and then comparing the average ranks to determine if there are statistically significant differences between the groups. The Friedman statistics τ_{χ^2} can be computed as follows:

$$\tau_{\chi^2} = \frac{12N}{k(k+1)}\left(\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right),$$

where k and N are the number of algorithms and datasets, respectively, and R_j is the sum of ranks for the j-th group.

The Friedman test can be followed by post-hoc tests (e.g., Nemenyi test, [28]) to identify which specific groups differ from each other when significant differences are found. The performance disparity between the two clustering methods is deemed significant when the average rank difference between them surpasses the critical threshold. The critical difference can be computed as follows:

$$CD = q_a\sqrt{\frac{k(k+1)}{6N}},$$

where q_a is the critical value for the Nemenyi test at a specific significance level (alpha), which can be found in Nemenyi's critical values table.

2.3. Data normalization

In a multi-indicator system, indicators often possess varying scales and quantitative levels due to their differing nature and meaning. Directly using raw information for an analysis can result in the attenuation of indicators with lower values and the overemphasis of those with higher values. Therefore, this section employs the normalization method to process the continuous indicators dimensionlessly, thus ensuring the reliability and validity of the analysis results. The calculation formula for normalization is as follows:

$$x_{ij}^* = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})},$$

where x^*_{ij} represents the normalized value, x_{ij} is the original measured value, $\min(x_{ij})$ denotes the minimum value of the variable, and $\max(x_{ij})$ denotes the maximum value of the variable.

2.4. Feature vectorization and selection

Categorical data is comprised of variable indicators that need to undergo vectorization prior to modeling and analysis. For instance, concerning the patient survival status, "0" signifies the patient's survival during their ICU stay, whereas "1" signifies the patient's demise during the ICU stay. Similarly, in terms of gender, "0" corresponds to male, while "1" corresponds to female. When dealing with categorical variables such as race, onehot coding was employed for the purpose of vectorization. This

approach facilitates the representation of categorical patient characteristics in a structured and vectorized format.

High correlations between variables can introduce bias into the model, which impacts the estimation of explanatory variables. This bias can lead to inconsistent results compared to actual outcomes and may even impede the model convergence. This study employed Pearson correlation coefficients and the recursive feature elimination (RFE) technique to select features from the training cohort data. In this context, variables with correlation coefficients that exceed 0.7 will be removed and processed accordingly.

2.5. Modelling

The study cohort exhibited a severe data imbalance issue, with the ratio of in-hospital survival to in-hospital mortality standing at 1150 : 116. To address this significant class imbalance, we have devised diverse methodologies, broadly categorized into two primary approaches: data-driven and algorithm-driven methods. Besides, hyperparameter optimization and cross-validation through GridSearchCV were applied to prevent overfitting and to increase model accuracy. Figure 2 shows the simplified schematic workflow for the predictive model.

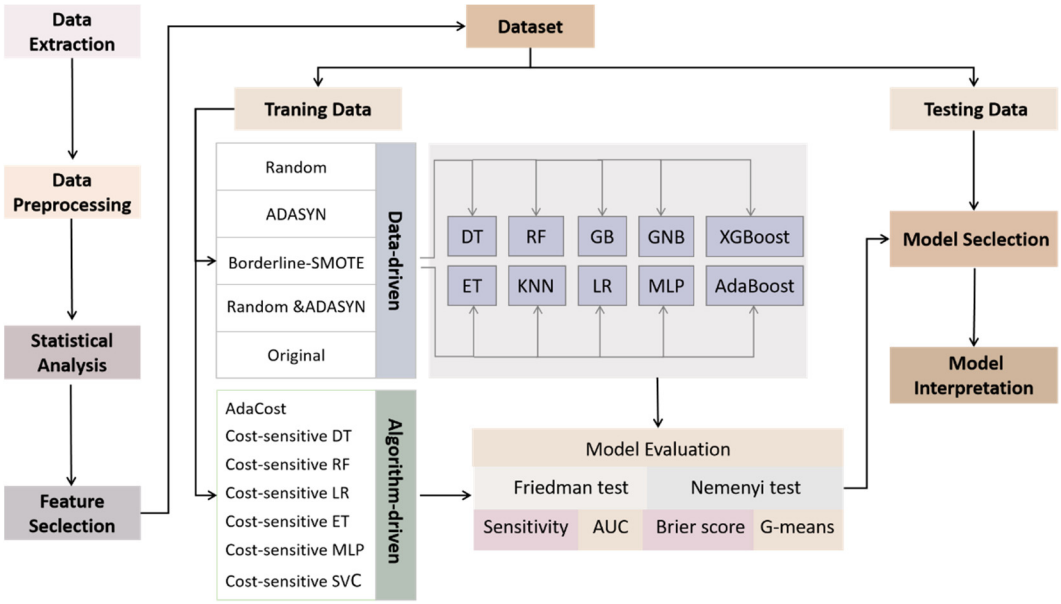

Figure 2. The simplified schematic workflow for the predictive model. It roughly includes the following steps: data filtering and preprocessing, feature selection, model optimization, and interpretation.

2.5.1. Data-driven approach

Data-driven approaches aim to balance class distribution in the input dataset by adjusting the class ratio. Commonly employed sampling techniques, such as under-sampling, oversampling, or a combination of both, have been widely utilized for this purpose [19].

In our study, we implemented four resampling methods: random (ROS) oversampling,

Borderline-SMOTE (BS) oversampling, adaptive synthetic (ADASYN) oversampling, and a combination of random under-sampling and ADASYN oversampling (ADASYN&ROS). The ROS technique is a data-level approach aimed at addressing the issue of imbalanced data by increasing the number of minority classes. This is accomplished by randomly replicating instances to balance the majority classes [29]. BS generates synthetic minority class samples by interpolating features from existing instances in this border region, and aims to improve the overall balance in the dataset for an enhanced ML model performance [30]. ADASYN generates synthetic minority class samples based on the number of neighboring samples for each data point, placing a greater emphasis on those minority class samples that are relatively isolated in the feature space [31]. ADASYN&ROS involves random under sampling of the majority class samples, specifically in-hospital survival data, concurrently with ADASYN oversampling of the minority class samples, corresponding to in-hospital mortality data. This dual approach aims to address class imbalance while preserving the diversity of the dataset.

Subsequently, ML algorithms are trained on the aforementioned four resampled training sets, and a model validation was performed on the corresponding test sets. The candidate ML algorithms employed in this study include a logistic regression (LR), random forest (RF), decision tree (DT), extra tree (ET), gradient boosting (GB), extreme gradient boosting (XGBoost), Gaussian Naive Bayes (GNB), multilayer perceptron (MLP), adaptive boosting (Adaboost), and k-nearest neighbors (KNN). The selection of these diverse algorithms is motivated by the unique advantages inherent in each technique. The optimal approach for the RHD dataset remains uncertain. Nevertheless, among the mentioned candidate methods, some have demonstrated success in simulating the prediction of in-hospital mortality for patients in the ICU [6], while others represent emerging methodologies in the field [10,14].

2.5.2. Algorithm-driven approach

On the other hand, the algorithm-driven approach seeks to fine-tune the learning algorithm or classifier without necessitating modifications to the original training dataset [32]. The basic assumption by any traditional classification algorithm is that the cost of misclassification is the same for all the response variable values. In this paper, we utilized the cost-sensitive learning methods by deliberately increasing the cost of misclassified samples and adjusting their weights during the training process.

To be specific, we implemented the adaptive cost-sensitive boosting (AdaCost) algorithm [33], which adapts sample weights in real-time based on their classification errors to mitigate the impact of high-cost classification errors. AdaCost achieves this by refining the weights of incorrectly classified samples in accordance with the predefined cost values present in the cost matrix. This results in a more substantial penalty for high-cost classification errors, thereby elevating the model's sensitivity to such errors during the training phase. This comprehensive approach places a heightened emphasis on addressing high-cost classification errors and, consequently, enhances the model's cost sensitivity. Furthermore, our research was extended to encompass a range of cost-sensitive learning techniques, including the cost-sensitive support vector classification (SVC), the cost-sensitive LR, the cost-sensitive DT, the cost-sensitive ET, the cost-sensitive MLP, and the cost-sensitive RF. These diverse methodologies collectively contribute to a holistic approach aimed at tackling data imbalance issues with a robust theoretical foundation.

2.5.3 Model evaluation

The assessment of the candidate models' performance encompasses various metrics, including the area under the receiver operating characteristic curve (AUC), sensitivity, Brier score, geometric mean score (G-means), and area under the precision-recall curve (PR-AUC). AUC acts as a valuable metric that captures the balance between the true positive rate and the false positive rate at different thresholds. It offers a quantitative evaluation of a model's proficiency in distinguishing between distinct classes. The Brier Score is a metric used to assess the quality of probability predictions, which is calculated as follows:

$$\text{Brier Score} = \frac{\sum (y - f_i)^2}{N},$$

where *N* represents the number of samples, *y* represents the actual class labels (0 or 1), and fi represents the model's probability predictions for the ith sample being positive. In scenarios characterized by imbalanced datasets, solely relying on traditional metrics like accuracy can be insufficient. This limitation arises from the model's tendency to favor the majority class, potentially resulting in a suboptimal performance for the minority class. Sensitivity measures the model's capability to accurately identify positive class samples, which is as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

where TP is the true positive count, and FN is the false negative count. The G-means score introduces a comprehensive approach to the performance evaluation, aiding in the identification and mitigation of such imbalances. It can be calculated as follows:

$$\text{G-mean} = \sqrt{Sensitivity \times Specificity},$$

where specificity is true negative rate. On the other hand, PR-AUC assesses a model's performance under varying trade-offs between precision and recall. They both are commonly employed to evaluate the model's effectiveness in handling imbalanced datasets. The values of the aforementioned metrics all lie within the range of 0 to 1. In the case of the Brier Score, a proximity to 0 signifies a superior model performance. Conversely, for the remaining metrics, a greater proximity to 1 indicates an improved model performance.

2.5.4. Model interpretation

SHAP is a tool to measure the variable importance, and understands the global model structure based on combining local explanations of each prediction [25]. Its method of evaluating feature importance is to use coalitions of one or more features to predict a quantity of interest, to find the difference in predictions between coalitions that do and do not include a given feature, and then to use such differences to quantify the magnitudes and directions of each feature's global and local contributions. The method works by constructing an additive explanation model:

$$g(z') = \phi_0 + \sum_{i=1}^{m} \phi_i z_i',$$

where g is the explanation model, $z \in (0, 1)^m$, m is the number of predictor variables, z_i' is a coalition vector, $\phi_i \in R$ is the importance value of the ith predictor, and ϕ_0 is the expected value of the target variable, or in other words, the mean of all prediction. For any predicted sample, the Shapley value Φ_i is as follows:

$$\Phi_i = \sum_{S \subseteq m \setminus i} \frac{|S|!(M-|S|-1)!}{M!}[f(S \cup i) - f(S)],$$

where $f(S \cup i)$ and $f(S)$ are the model outcomes with and without the ith predictor, respectively, and S is a subset of features used in the model. Values of Φ_i that are greater (less) than zero refer to the positive (negative) effect of the variable i, which increases (decreases) the predicted value above (below) the base value. In our study, we employed a SHAP feature importance assessment to provide a comprehensive global interpretation of the baseline model we developed. Furthermore, we utilized SHAP to generate illustrative instances that demonstrate the local explanations for individual predictions.

Local interpretable model-agnostic explanations (LIME) is an interpretation method for local and individual model explanations, and is designed to aid in understanding the reasons behind ML model predictions for specific instances [26]. The core idea of this approach is to approximate the behavior of the original model near a particular instance by constructing a simple and interpretable local model. The formula for LIME is represented as follows. For the original model f and the instance to be explained x, LIME constructs a local model g to approximate f around x. The representation of g is given by the following:

$$g(z) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g),$$

where $L(f, g, \pi x)$ represents the loss function, which measures the disparity in predictions between g and f on the perturbed samples, $\Omega(g)$ is the regularization term, which ensures the simplicity of g, and πx is the weight function used to assign weights to perturbed samples, which is usually determined based on the similarity between the instance and the perturbed samples. Figure 2 shows the simplified schematic workflow of this study.

3. Results

After calculating the Pearson correlation matrix for 67 indicators, we identified and subsequently removed 20 variables that exhibited multicollinearity. Following a specific feature ranking criterion, RFE initiates with a complete set and iteratively eliminates the least relevant features, resulting in the selection of the top 26 important features. Figure 3 illustrates the Pearson correlation matrix among the remaining 26 indicators. All the correlation coefficients are less than 0.7, which addresses the potential issue of multicollinearity.

From a data-driven perspective, we reconstructed the dataset through oversampling techniques to mitigate the negative impact of sample skewness during the learning process. Specifically, we explored four resampling methods (ROS, BS, ADASYN, ROS-ADASYN) in combination with ten ML classifiers to predict the death risk of the test cohort. Figures 4(a)–(e) depict the class distribution of "Survive" and "Death" in the original dataset and in resampled datasets, specifically focusing on two key features, namely wbc_max and sofa. In the original dataset, the class ratio is observed at 805 : 81. Subsequently, the data was subjected to diverse resampling techniques, resulting in class ratios of 805 :

805, 805 : 805, 922 : 919, and 805:805, respectively (Figure 4(f)).

Figure 5(a) presents the predictive performance for ten ML models without a resampling technique (e.g., on the original dataset). For convenience, we referred to these models as O-models in the following. It is conspicuously evident that, prior to implementing data balancing techniques, the O-models exhibited a discernible trend of lower sensitivity and G-means values across the remaining models. This occurrence can be attributed to the inherent nature of a direct analysis on severely imbalanced datasets, which results in a pronounced bias within the trained models towards the majority class; in this context, it pertains to the survival outcome. While the O-models boasted a commendable AUC and Brier score, they were afflicted by a significantly low true positive rate. This, in practice, rendered it ill-suited for the task of effectively distinguishing the outcome of deceased patients, thus diminishing its practical reliability.

Figure 5(b)–(e) provides a comprehensive visual representation of the predictive performance of various combinations of four re-sampling algorithms paired with ten ML models. After the application of data balancing strategies, these re-sampling combined models exhibited consistently favorable AUCs, and retained promising Brier scores. It is noteworthy that virtually all re-sampling combined models showcased significant enhancements in both the sensitivity and G-means. This enhancement was most pronounced in models such as ROS-LR, BS-LR, ADASYN-LR, and ADASYN&ROSLR, which signifies a notable augmentation in the ability of the trained models to discriminate between positive and negative samples.

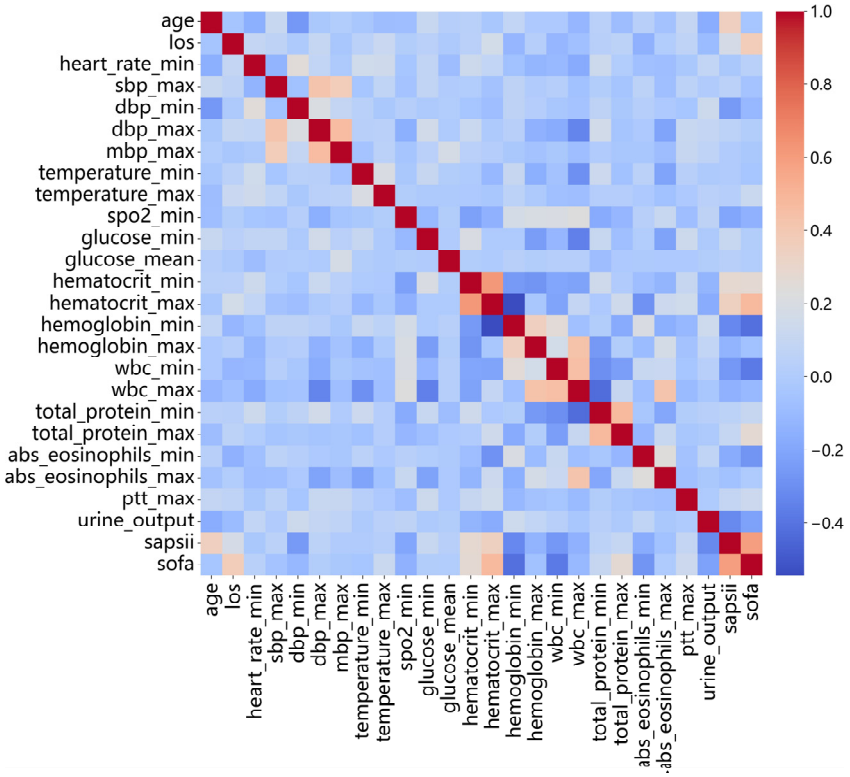

Figure 3. The heatmap of Pearson correlation coefficients among 26 selected features.

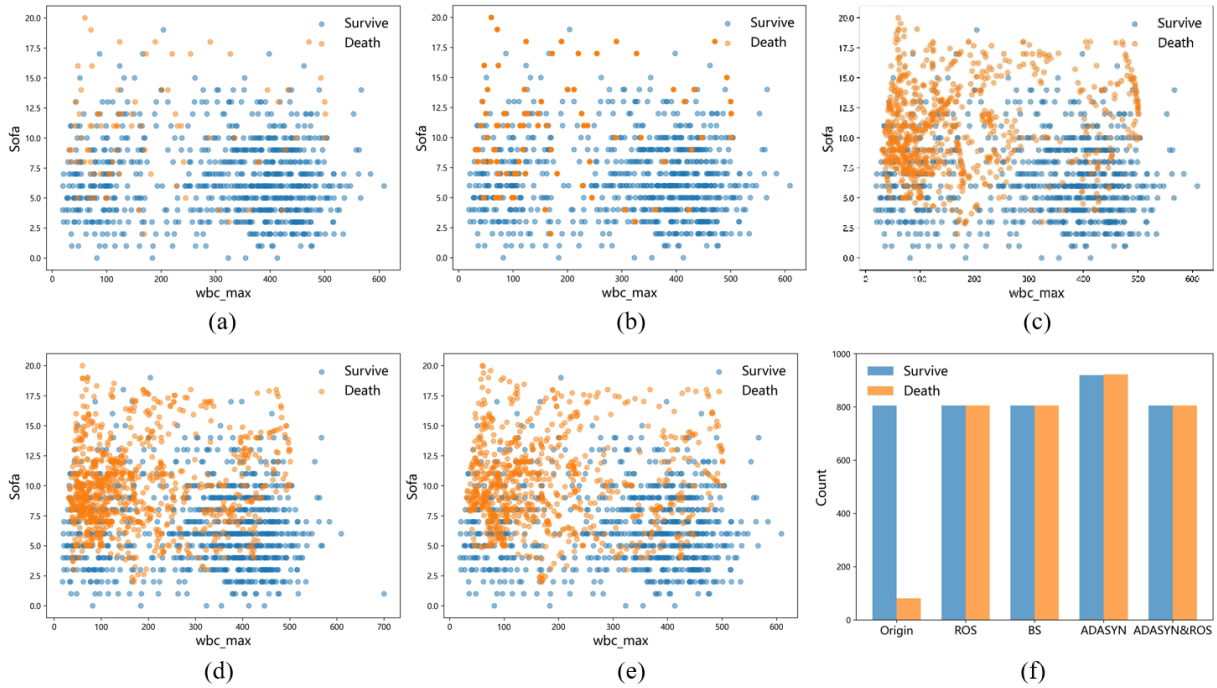

Figure 4. (a)–(e) The class distribution of "Survive" and "Death" in the original dataset, as well as in datasets resampled using ROS, BS, ADASYN, and ROS-ADASYN techniques. (f) The counts of "Survive" and "Death" for each of the five datasets.

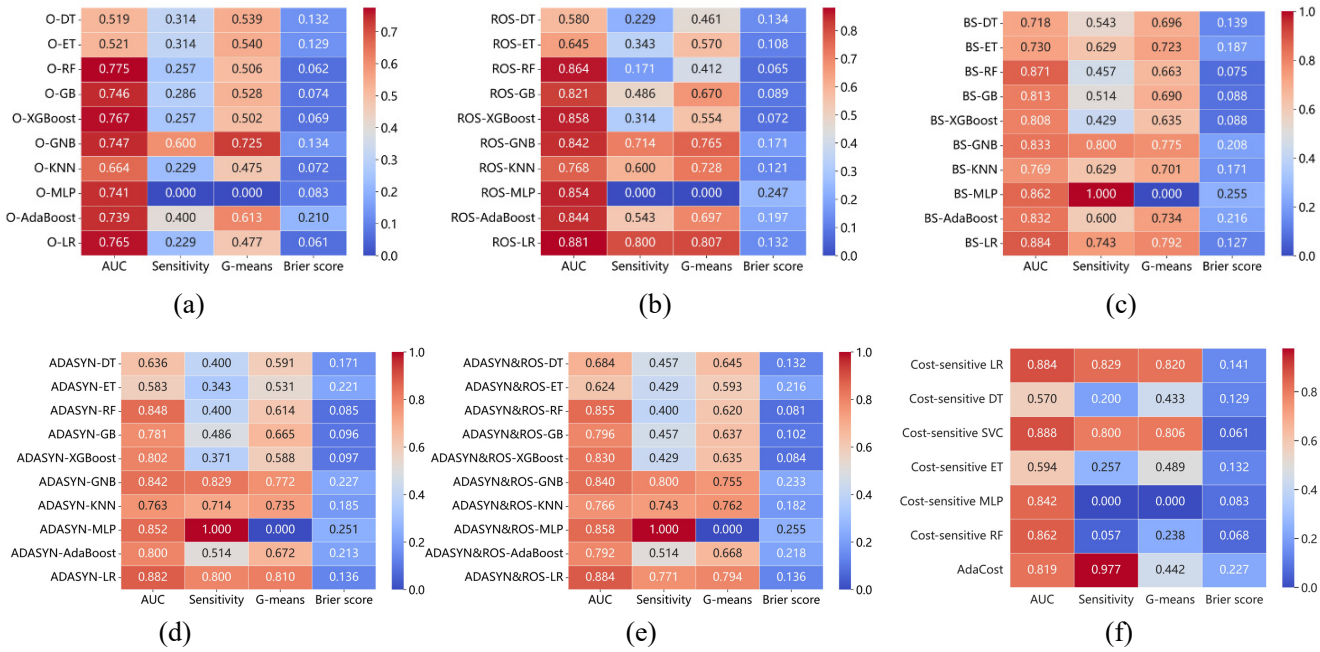

Figure 5. (a)–(e) Evaluated metrics for 10 ML models on the original, ROS resampled, BS resampled, ADASYN resampled and ADASYN&ROS datasets, respectively. (f) Evaluated metrics for seven cost-sensitive ML models.

Figure 5(f) presents a comprehensive evaluation of seven cost-sensitive ML models. The outcomes of this evaluation underscore the efficacy of cost-sensitive learning techniques in improving the model performance. This is achieved by assigning higher rewards to the minority samples, thereby increasing the accuracy of classifying these minority instances. Notably, the cost-sensitive SVC stands out as the most distinguished method in this context. It exhibits the highest discrimination levels, as evidenced by an AUC of 0.888 and a Brier score of 0.061. Additionally, this model demonstrates a commendable sensitivity of 0.800 and an impressive G-mean of 0.806. Given its exceptional performance, the cost-sensitive SVC appears to be a suitable base model for an interpretable analysis. Subsequently, we conducted a Friedman test to further assess the performance of these seven ML models. We computed the average rank for each algorithm across the entire dataset. In our analysis, using a significance level of $\alpha = 0.05$, the test statistics for the AUC, sensitivity, G-means, and Brier score yielded values of 19.689, 202.067, 78.918, and 138.860, respectively. All of these values surpass the critical threshold of 2.324, indicating statistically significant distinctions among these methods. Subsequently, we performed a post-hoc analysis using the Nemenyi test to discern the nature of these differences. In our scenario, the CD was determined to be 3.185. Figure 6 illustrates the CD diagram derived from pairwise Nemenyi tests. When the black lines connect certain groups in the figure, it indicates that the differences between these groups are not significant in multiple comparisons. To further offer a holistic insight into the predictive performance of the aforementioned 57 ML models, we present their PR-AUC and AUC curves in Figures 7 and 8, respectively. Notably, one striking revelation is the exceptionally high PR-AUC of 0.977 achieved by the Adacost model. It also stands as a favorable choice in practical applications, where the emphasis is on the identification of mortality. Here, we are particularly focused on the overall performance across four key metrics, and therefore, the cost-sensitive SVC should be the preferred method for practical clinical diagnoses and applications.

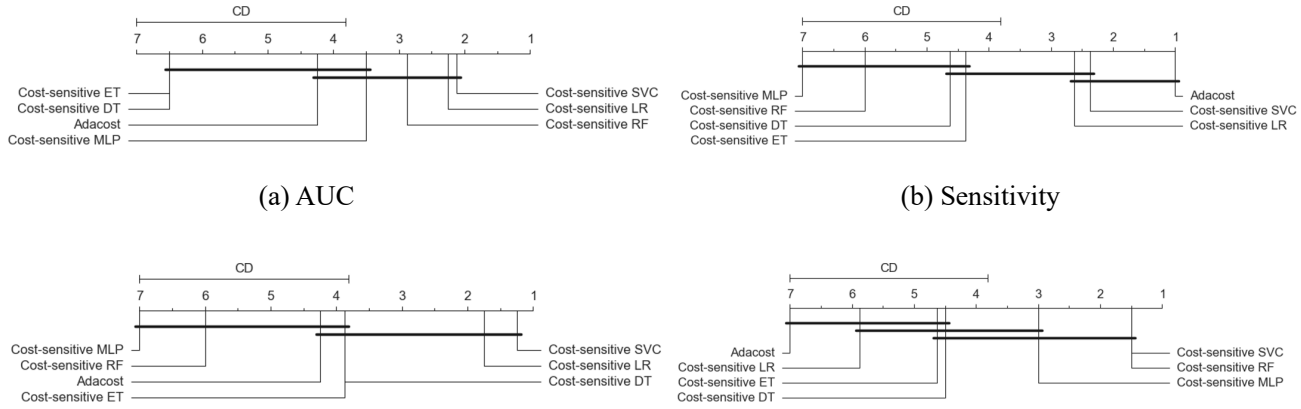

Figure 6. The CD diagram of AUC, sensitivity, G-means, and Brier score derived from pairwise Nemenyi tests for the algorithm-driven models.

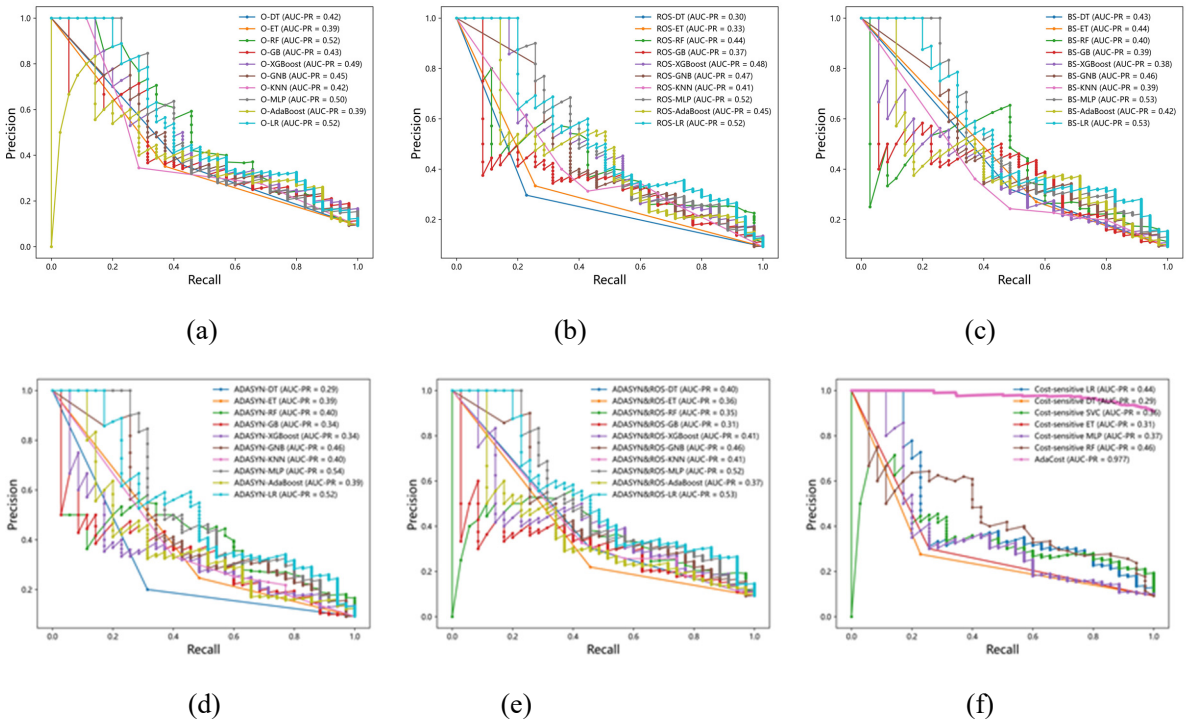

Figure 7. The PR-AUC curves of data and algorithm-driven approaches for the test cohort.

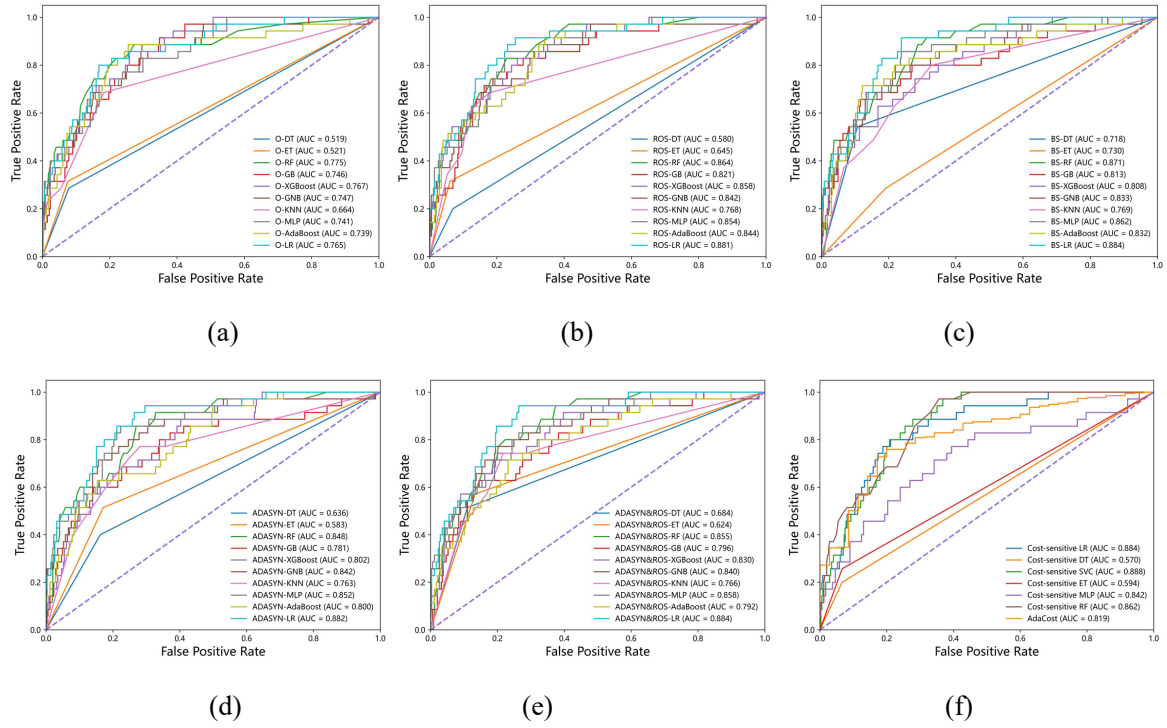

Figure 8. The AUC curves of data and algorithm-driven approaches for the test cohort.

In SHAP, the global importance of each feature we estimated was used to understand the general impact of various features across the study cohort (Figure 9). The SHAP summary plot illustrated the

entire distribution of each feature's impact on the model output. The color allowed us to understand how changes in the value of a feature affected the change in outcome. Red represents a high feature value, whereas blue represents a low feature value. The further away a point is from the baseline SHAP value of zero, the stronger it effects the output. This way, a features relationship with the SHAP value (and in turn the predicted output) can be better understood.

Our findings reveal that the physiological parameters (sofa, age, total_protein_max) and immune function indicators (wbc_max, wbc_min), in conjunction with markers reflecting circulatory dysfunction (dbp_max, temperature_max, urine output), and coagulation function (ptt_max), emerge as pivotal determinants to evaluate the mortality risk in 57 patients upon ICU admission. These factors may exert a more pronounced influence on assessing the patient's condition and prognosis compared to other risk factors, including the blood glucose and hematocrit levels. The directional impact of the SHAP values suggests that a right-tailed distribution of elevated sofa, age, body temperature, white blood cell count, and partial thromboplastin time is associated with an increased risk of mortality. Additionally, a left-tailed distribution of low urine output, total protein level, and maximum diastolic blood pressure shows significant associations and exhibits a positive correlation with the predicted mortality.

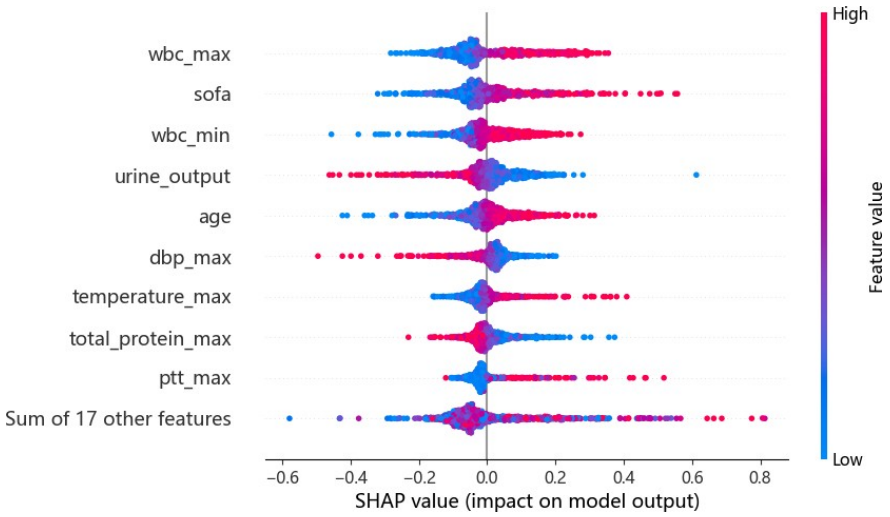

Figure 9. SHAP summary plot of the cost-sensitive SVC model. Each dot represents a patient's feature attribution value. Dots are color-coded based on patient feature values, with red indicating higher values (e.g., higher risk of death) and blue indicating lower values. The distance from the baseline SHAP value of zero reflects the strength of each feature's impact on the model output.

In Figure 10, we showcase the SHAP interactive dependency plot. This visualization, which portrays the concurrent variation of two features and its consequential impact on the SHAP values and the ultimate model output, serves as a valuable tool to discern the sensitivity of predictions to various features. In Figure 10(a), a noticeable trend emerges: as the sofa values rise, there is a concurrent increase in the corresponding SHAP values. This relationship is easily interpretable, as an escalation in the sofa score typically aligns with a deterioration in a patient's condition and organ functionality, ultimately heightening the risk of mortality. When the age surpasses 72.5 years, the associated SHAP values consistently maintain positivity, thus exhibiting a gradual increase with advancing age. This pattern signifies a persistent escalation in the risk of mortality. Notably, an interaction between the sofa

score and age emerges during this period. Among patients of an identical age, those with elevated sofa scores exhibit a heightened susceptibility to mortality. Contrastingly, this interaction is not discernible below the age of 72.5 years (Figure 10(b)). With the extension of the maximum partial thromboplastin time (ptt max), the corresponding SHAP values shift from negative to positive, thus increasing the risk of mortality in RHD patients. This indicates potential abnormalities in the coagulation function associated with prolonged ptt max, thereby elevating the risk of thrombotic events or other coagulation-related complications in RHD patients.

When ppt_max exceeds 120 seconds, there may be an interaction between ppt_max and sofa: for cases with the same ppt_max value, when the sofa is higher, there are higher SHAP values, indicating an increased risk of mortality (Figure 10(b)). For urine output, dbp_max, and total_protein_max, as their values increase, the corresponding SHAP values also turn positive, indicating a decrease in the associated risk of mortality. This phenomenon can be attributed to the body's improvement in the corresponding physiological conditions. Specifically, an elevated urine output suggests an enhanced physiological circulation and renal function, an increase in dbp_max reflects an improved diastolic blood pressure regulation, and a rise in total_protein_max signifies positive changes in the overall protein levels. These improvements collectively contribute to a more favorable health status and a reduced risk of mortality (Figures 10(d)–(f)).

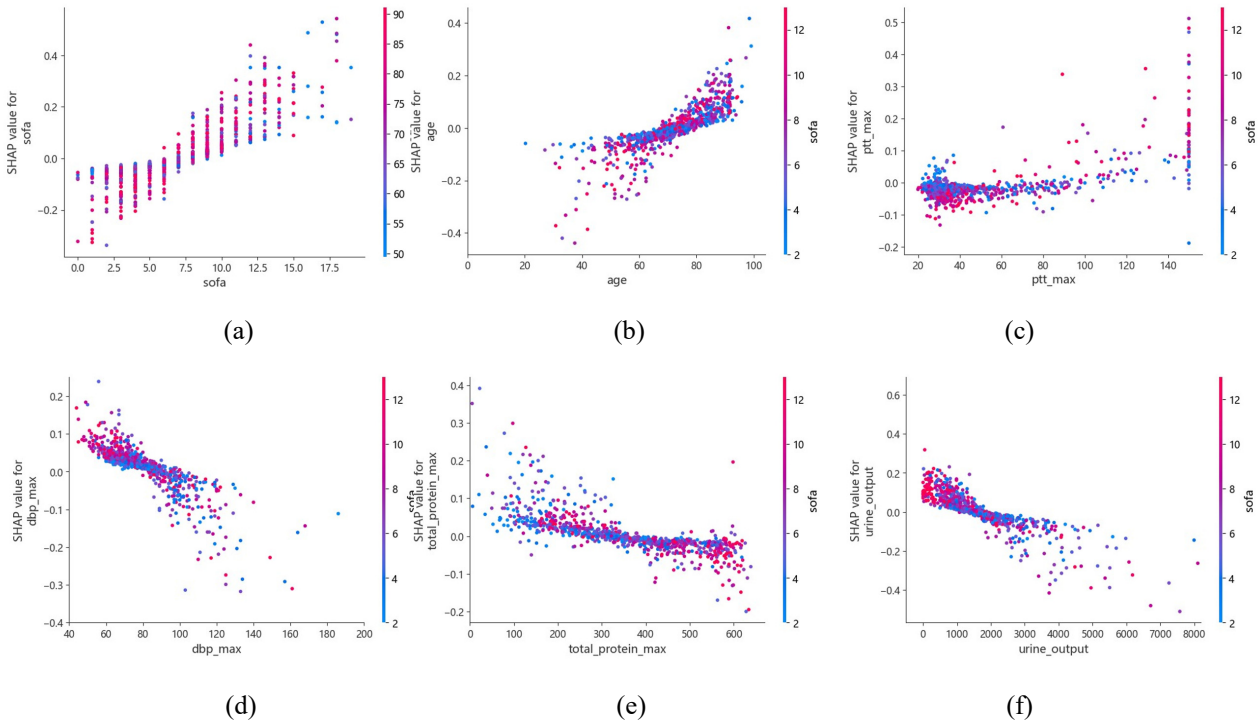

Figure 10. SHAP dependence plot of the cost-sensitive SVC model. The x-axis denotes the value of the primary driver, while the y-axis represents the corresponding SHAP value. The coloration of each point signifies the value associated with the secondary driver.

We randomly selected two positive cases (deaths) from the test dataset and conducted a local analysis of the mortality risk using both SHAP and LIME interpreters (Figure 11). It is evident that the top 10 key features identified by these two interpreters are remarkably similar, and the response trends

of these features to the risk of death are consistent.

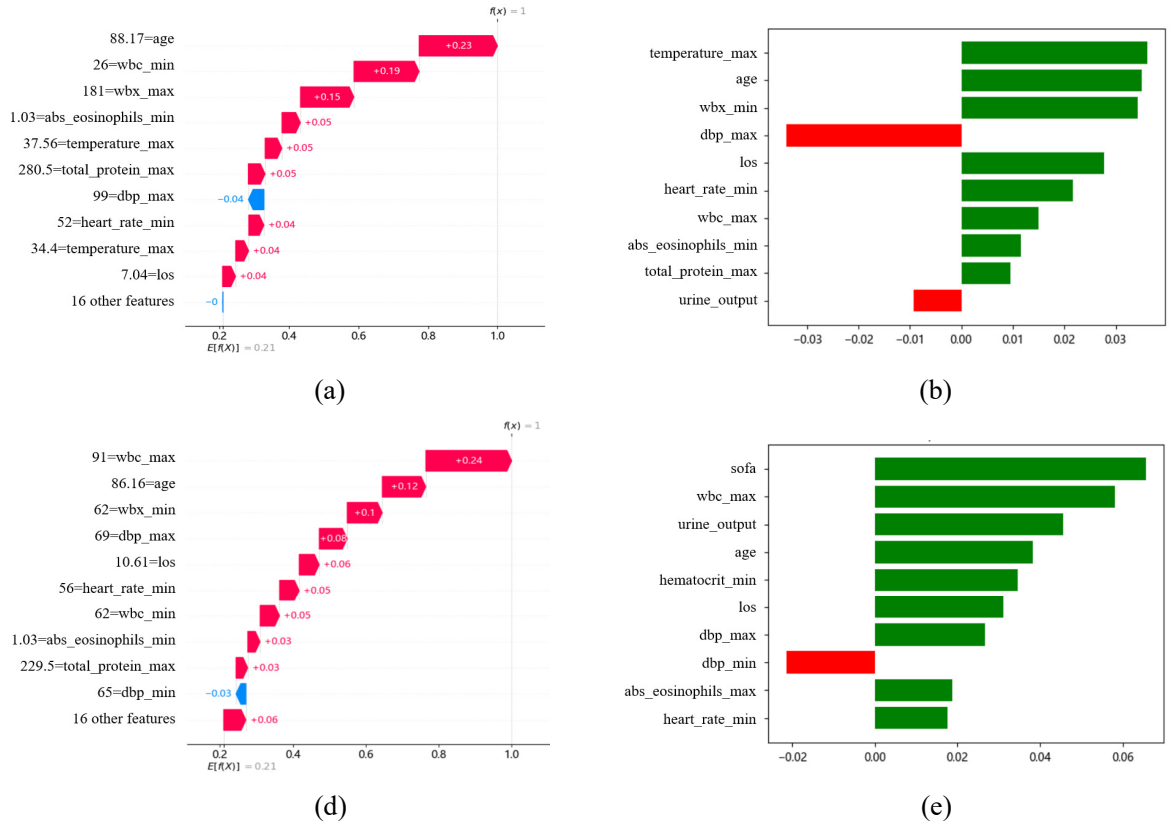

Figure 11. Local interpretable of the model's true positive cases (deaths). (a) and (c): True positive individual interpretation based on SHAP. (b) and (d): True positive individual interpretation based on LIME.

4. Discussion

RHD, as a disease associated with poverty, continues to impose a disproportionately high burden on the developing world, despite being fundamentally preventable. Compared to other prevalent infectious issues such as malaria, HIV/AIDS, and tuberculosis, RHD receives relatively less attention in the medical and scientific communities, despite causing a significant cardiovascular morbidity and mortality burden. This relative neglect and lack of funding may have contributed to the limited progress in basic medical research in this field over the past five decades [34]. Specifically, the current research landscape regarding accurately predicting the mortality risk in RHD remains relatively underexplored.

By systematically combining four resampling methods (ROS, BS, ADASYN, ROS-ADASYN) with ten ML techniques, along with considering seven cost-sensitive learning models, we presented a total of 57 ML regression prediction models in this study. We identified that the utilization of the cost-sensitive SVC yielded the most favorable outcomes (with an AUC of 0.888, sensitivity of 0.800, G-means of 0.806, and a Brier score of 0.061), which was well-suited for our research cohort. Given that the negative class (deaths) is often the minority class, classifiers without balancing practices may exhibit a bias toward predicting the majority class. We demonstrated that the application of re-sampling and cost-sensitive learning methods significantly enhanced the sensitivity and G-means for the

minority class. Although the improvement in the AUC was not statistically significant, resampling and cost-sensitive learning techniques contributed to balancing the weights between high-risk and low-risk patients. Integrating resampling and cost-sensitive learning techniques into mortality risk prediction models should be more widely applied in real clinical cases to address the challenges posed by imbalanced data.

Global interpretability enables clinical physicians to understand the response trends of the model across the entire feature space. In contrast, local interpretability provides feature-based decision explanations for specific individuals. In practice, both these approaches can assist clinical physicians in making effective decisions during medical processes. In our study, mortality predictors of RHD patients were examined and were found to be consistent with previous findings. Our observation of increased the mortality rates in older individuals with RHD, which is in line with findings from earlier studies [35,36].

One main contributing factor to this trend is the increased incidence of two high fatality complications of RHD, namely atrial fibrillation and stroke, which rise with the age at diagnosis (by 5% and 4% per year of age, respectively) [35]. PTT emerged as a pivotal predictor within the cost-sensitive SVC regression model. Heightened levels of fibrinogen correlated with the presence of cerebral microbleeds in ischemic stroke patients afflicted by RHD [37]. Anticoagulation stood as the cornerstone for mitigating the progression of RHD, especially when evaluating coagulation function indices such as PT. This is particularly crucial for patients with RHD, especially those grappling with atrial fibrillation, a history of thromboembolism, or left atrial thrombosis [38]. In patients with RHD, the heart is susceptible to the effects of rheumatic inflammation, leading to valve damage and structural alterations. These changes can impact both the contraction and relaxation functions of the heart, which subsequently influences the diastolic pressure. A low diastolic pressure may result in a decline in systemic tissue perfusion, which negatively affects the normal function of organs. During cardiac diastole, a low diastolic pressure may reduce the filling of the coronary arteries, causing an insufficient blood supply to the heart. Simultaneously, it may also decrease the blood supply to other organs throughout the body. In the lungs, a low diastolic pressure might induce congestion in the pulmonary circulation, leading to pulmonary congestion. The deterioration of the left ventricular (LV) diastolic function, which causes an increase in LV filling pressure and pulmonary congestion, could potentially trigger acute heart failure (AHF). Severe acute ischemia rapidly impairs myocardial relaxation, which affects early LV filling and further elevates the filling pressure. In cases of sudden onset atrial fibrillation with the loss of atrial contraction, LV filling may be compromised, which significantly increases the filling pressure in the presence of a pre-existing diastolic dysfunction. For instance, severe mitral valve stenosis (common in RHD) represents a diastolic dysfunction caused by valve abnormalities rather than LV structural disease. Moreover, it can precipitate atrial fibrillation, which further escalates the risk of AHF. An appropriate diastolic pressure level signifies normal perfusion to the heart and organs throughout the body, which serves as a vital indicator to assess the risk of complications such as AHF and atrial fibrillation [29]. The case report and experimental analysis [39–41] confirmed that RHD adversely affects cardiac function, potentially resulting in a decreased cardiac pumping capacity and reduced tissue perfusion throughout the body. This prompts the kidneys to sense the decreased blood flow, which leads to a reduction in urine production. However, a significant improvement in the condition is observed after administering cardiac and diuretic medications. This improvement can be attributed to the alleviation of pulmonary vein congestion through treatment, which prolongs the diastolic time and enhances the cardiac output [9]. The sofa score serves as a straightforward, yet potent,

rating index. This score quantifies organ damage by measuring the burden of organ dysfunction in critically ill individuals, and encompasses assessments of cardiovascular, hemostatic, and renal functions [42]. Besides, total protein is a reliable indicator of nutritional status [43], and there is a report linking malnutrition to an elevated mortality in individuals undergoing cardiac surgery [44]. Insufficient total protein levels may lead to inadequate blood volume and fluid retention, potentially exacerbating symptoms of heart failure. Previous investigations have indicated that hypoalbuminemia was common in patients with stable chronic or acute heart failure [45,46]. Moreover, it can impact the transport capacity of blood, thus influencing organ perfusion and oxygenation. Moreover, reduced total protein levels may compromise immune function, thus increasing the risk of infections; from this, the catabolism will increase, which includes reduced protein synthesis rates and increased protein degradation rates [47]. Therefore, when assessing the risk of mortality in patients with RHD, consideration of total protein levels is deemed a critical factor. Therefore, monitoring changes in the urine output, the sofa score, and the total protein levels is pivotal to evaluate the patient conditions and to assess the mortality risk.

5. Conclusions

In response to the existing issues of data imbalance and a lack of interpretability in predicting the mortality risk of RHD patients in the ICU, we have developed a comprehensive workflow. This workflow integrated feature selection, resampling techniques, cost-sensitive learning methods, the Friedman test, the Nemenyi test, and interpretable ML approaches. The key findings include the following:

(1) The cost-sensitive SVC model demonstrated a superior performance among the 57 predictive models constructed;

(2) Both resampling and cost-sensitive learning methods resulted in significant improvements compared to models directly built on the original data in the predictive performance, especially in terms of specificity and G means; and

(3) Unveiling key physiological features and their response trends to RHD-related mortality.

ML is a valuable and increasingly necessary tool in modern healthcare systems. The ML models developed in this study have undergone rigorous validation and demonstrated state-of-the-art performance. Their key advantage lies in their ability to be easily interpreted by clinical professionals solely using preoperative data. Accurately quantifying the risk of postoperative mortality can better inform patient-centered decision-making. Additionally, it can guide targeted quality improvement interventions and support activities of accountable care organizations relying on precise population risk estimates. Future work will focus on assessing the effectiveness and efficiency of integrating model predictions into clinical decision processes and improving hospital costs.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by National Natural Science Foundation of China (12201577, U23A2065), Key Scientific and Technological Project of Henan Province (232102320136).

Conflict of interest

The authors declare no conflict of interest.

References

1. Marijon E, Mirabel M, Celermajer DS, Jouven X, (2012) Rheumatic heart disease. *Lancet* 379: 953–964. https://doi.org/10.1016/S0140-6736(11)61171-9
2. Carapetis JR, Beaton A, Cunningham MW, Guilherme L, Karthikeyan G, Mayosi BM, et al. (2016) Acute rheumatic fever and rheumatic heart disease. *Nat Rev Dis Primers* 2: 15084. https://doi.org/10.1038/nrdp.2015.84
3. Muhamed B, Parks T, Sliwa K, (2020) Genetics of rheumatic fever and rheumatic heart disease. *Nat Rev Cardiol* 17: 145–154. https://doi.org/10.1038/s41569-019-0258-2
4. Ordunez P, Martinez R, Soliz P, Giraldo G, Mujica OJ, Nordet P, et al. (2019) Rheumatic heart disease burden, trends, and inequalities in the Americas, 1990–2017: A population-based study. *Lancet Global Health* 7: e1388–e1397. https://doi.org/10.1016/S2214-109X(19)30360-2
5. Watkins DA, Johnson CO, Colquhoun SM, Karthikeyan G, Beaton A, Bukhman G, et al. (2017) Global, regional, and national burden of rheumatic heart disease, 1990–2015. *N Engl J Med* 377: 713–722. https://doi.org/10.1056/NEJMoa1603693
6. Xie J, Su B, Li C, Lin K, Li H, Hu Y, et al. (2017) A review of modeling methods for predicting in-hospital mortality of patients in intensive care unit. *J Emerg Crit Care Med* 1: 1–10.
7. Rehman S, Akhtar N, Saba N, Munir S, Ahmed W, Mohyuddin A, et al. (2013) A study on the association of TNF-α-308, IL-6-174, IL-10-1082 and IL-1RaVNTR gene polymorphisms with rheumatic heart disease in Pakistani patients. *Cytokine* 61: 527–531. https://doi.org/10.1016/j.cyto.2012.10.020
8. Dooley LM, Ahmad TB, Pandey M, Good MF, Kotiw M, (2021) Rheumatic heart disease: A review of the current status of global research activity. *Autoimmun Rev* 20: 102740. https://doi.org/10.1016/j.autrev.2020.102740
9. Arvind B, Ramakrishnan S, (2020) Rheumatic fever and rheumatic heart disease in children. *Indian J Pediatr* 87: 305–311. https://doi.org/10.1007/s12098-019-03128-7
10. Xu Y, Han D, Huang T, Zhang X, Lu H, Shen S, et al. (2022) Predicting ICU mortality in rheumatic heart disease: Comparison of XGBoost and logistic regression, 9: 847206. https://doi.org/10.3389/fcvm.2022.847206
11. Lee YW, Choi JW, Shin EH, (2021) Machine learning model for predicting malaria using clinical information. *Comput Biol Med* 129: 104151. https://doi.org/10.1016/j.compbiomed.2020.104151
12. Akter S, Das D, Haque RU, Tonmoy MIQ, Hasan MR, Mahjabeen S, et al. (2022) AD-CovNet: An exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with COVID-19. *Comput Biol Med* 146: 105657. https://doi.org/10.1016/j.compbiomed.2022.105657

13. Fan Z, Jiang J, Xiao C, Chen Y, Xia Q, Wang J, et al. (2023) Construction and validation of prognostic models in critically Ill patients with sepsis-associated acute kidney injury: Interpretable machine learning approach. *J Transl Med* 21: 406. https://doi.org/10.1186/s12967-023-04205-4
14. Martins JFB, Nascimento ER, Nascimento BR, Sable CA, Beaton AZ, Ribeiro AL, et al. (2021) Towards automatic diagnosis of rheumatic heart disease on echocardiographic exams through video-based deep learning. *J Am Med Inf Assoc* 28: 1834–1842. https://doi.org/10.1093/jamia/ocab061
15. Ali F, Hasan B, Ahmad H, Hoodbhoy Z, Bhuriwala Z, Hanif M, et al. (2021) Protocol: Detection of subclinical rheumatic heart disease in children using a deep learning algorithm on digital stethoscope: A study protocol. *BMJ Open* 11: e044070. https://doi.org/10.1136/bmjopen-2020-044070
16. Katarya R, Meena SK, (2021) Machine learning techniques for heart disease prediction: A comparative study and analysis. *Health Technol* 11: 87–97. https://doi.org/10.1007/s12553-020-00505-7
17. Shahid S, Khurram H, Billah B, Akbar A, Shehzad MA, Shabbir MF, (2022) Machine learning methods for predicting major types of rheumatic heart diseases in children of Southern Punjab, Pakistan. *Front. Cardiovasc. Med* 9: 996225. https://doi.org/10.3389/fcvm.2022.996225
18. Thabtah F, Hammoud S, Kamalov F, Gonsalves A, (2020) Data imbalance in classification: Experimental evaluation. *Inf Sci* 513: 429–441. https://doi.org/10.1016/j.ins.2019.11.004
19. Ghorbani M, Kazi A, Baghshah MS, Rabiee HR, Navab N, (2022) RA-GCN: Graph convolutional network for disease prediction problems with imbalanced data. *Med Image Anal* 75: 102272. https://doi.org/10.1016/j.media.2021.102272
20. Razzaghi T, Safro I, Ewing J, Sadrfaridpour E, Scott JD, (2019) Predictive models for bariatric surgery risks with imbalanced medical datasets. *Ann Oper Res* 280: 1–18. https://doi.org/10.1007/s10479-019-03156-8
21. Pera M, Gibert J, Gimeno M, Garsot E, Eizaguirre E, Miró M, et al. (2022) Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer. *Ann Surgery* 276: 776–783. https://doi.org/10.1097/SLA.0000000000005616
22. Ghamari SH, Abbasi-Kangevari M, Saeedi Moghaddam, S, Aminorroaya A, Rezaei N, Shobeiri P, et al. (2022) Rheumatic heart disease is a neglected disease relative to its burden worldwide: Findings from global burden of disease 2019. *J Am Heart Association* 11: e025284. https://doi.org/10.1161/JAHA.122.025284
23. Tao Y, Zhao J, Cui H, Liu L, He L, (2024) Exploring the impact of socioeconomic and natural factors on pulmonary tuberculosis incidence in China (2013–2019) using explainable machine learning: A nationwide study. *Acta Trop* 253: 107176. https://doi.org/10.1016/j.actatropica.2024.107176
24. Lundberg SM, Lee SI, (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Proc Syst* 2017: 30.
25. Ribeiro MT, Singh S, Guestrin C, (2016) "Why should I trust you?" Explaining the predictions of any classifier, In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144. https://doi.org/10.1145/2939672.2939778
26. Friedman M, (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 32: 675–701.
27. Nemenyi PB, (1963) *Distribution-Free Multiple Comparisons*, Princeton University.

28. Sharma S, Bellinger C, Krawczyk B, Zaiane O, Japkowicz N, (2018) Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance, In: *2018 IEEE International Conference on Data Mining (ICDM)*, 447–456. https://doi.org/10.1109/ICDM.2018.00060
29. Han H, Wang WY, Mao BH, (2005) Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, In: *International Conference on Intelligent Computing*, 878–887. https://doi.org/10.1007/11538059_91
30. He H, Bai Y, Garcia EA, Li S, (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning, In: *2008 IEEE International Joint Conference on Neural Networks*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969
31. Chawla NV, (2010) Data mining for imbalanced datasets: An overview, In: Maimon O, Rokach L, (eds) *Data Mining and Knowledge Discovery Handbook*, Boston: Springer, 565–601. https://doi.org/10.1007/978-0-387-09823-4_45
32. Fan W, Stolfo SJ, Zhang J, Chan PK, (1999) AdaCost: misclassification cost-sensitive boosting, In: *Proceeding of 16th International Conference on Machine Learning*, 99: 97–105.
33. Marijon E, Mocumbi A, Narayanan K, Jouven X, Celermajer DS, (2021) Persisting burden and challenges of rheumatic heart disease, *Eur Heart J* 42: 3338–3348. https://doi.org/10.1093/eurheartj/ehab407
34. He VY, Condon JR, Ralph AP, Zhao Y, Roberts K, de Dassel JL, et al. (2016) Long-term outcomes from acute rheumatic fever and rheumatic heart disease: A data-linkage and survival analysis approach. *Circulation* 134: 222–232. https://doi.org/10.1161/CIRCULATIONAHA.115.020966
35. Lawrence JG, Carapetis JR, Griffiths K, Edwards K, Condon JR, (2013) Acute rheumatic fever and rheumatic heart disease: Incidence and progression in the Northern Territory of Australia, 1997 to 2010. *Circulation*, 128: 492–501. https://doi.org/10.1161/CIRCULATIONAHA.113.001477
36. Liu J, Wang D, Xiong Y, Liu B, Lin J, Zhang S, et al. (2017) Association between coagulation function and cerebral microbleeds in ischemic stroke patients with atrial fibrillation and/or rheumatic heart disease. *Aging Dis* 8: 131. https://doi.org/10.14336%2FAD.2016.0715
37. Arrigo M, Jessup M, Mullens W, Reza N, Shah AM, Sliwa K, et al. (2020) Acute heart failure. *Nat Rev Dis Primers* 6: 16. https://doi.org/10.1038/s41572-020-0151-7
38. Pradhan RR, Jha A, Nepal G, Sharma M, (2018) Rheumatic heart disease with multiple systemic emboli: A rare occurrence in a single subject. *Cureus* 10: 7. https://doi.org/10.7759%2Fcureus.2964
39. DeBakey ME (1971) Left ventricular bypass pump for cardiac assistance: clinical experience. *Am J Cardiol* 27: 3–11. https://doi.org/10.1016/0002-9149(71)90076-2
40. Mickerson J, Swale J, (1959) Diuretic effect of steroid therapy in obstinate heart failure. *Br Med J* 1: 876. https://doi.org/10.1136%2Fbmj.1.5126.876
41. Janssens U, Dujardin R, Graf J, Lepper W, Ortlepp J, Merx M, et al. (2001) Value of SOFA (Sequential Organ Failure Assessment) score and total maximum SOFA score in 812 patients with acute cardiovascular disorders. *Crit Care* 5: 1. https://doi.org/10.1186/cc1292
42. McClave SA, Snider HL, Spain DA, (1999) Preoperative issues in clinical nutrition. *Chest* 115: 64S–70S. https://doi.org/10.1378/chest.115.suppl_2.64S
43. Evans AS, Hosseinian L, Mohabir T, Kurtis S, Mechanick JI, (2015) Nutrition and the cardiac surgery intensive care unit patient—An update. *J Cardiothorac Vasc Anesth* 29: 1044–1050. https://doi.org/10.1053/j.jvca.2015.03.021

44. Horwich TB, Kalantar-Zadeh K, MacLellan RW, Fonarow GC, (2008) Albumin levels predict survival in patients with systolic heart failure. *Am Heart J* 155: 883–889. https://doi.org/10.1016/j.ahj.2007.11.043
45. Uthamalingam S, Kandala J, Daley M, Patvardhan E, Capodilupo R, Moore SA, et al. (2010) Serum albumin and mortality in acutely decompensated heart failure. *Am Heart J* 160: 1149–1155. https://doi.org/10.1016/j.ahj.2010.09.004
46. Don BR, Kaysen G, (2004) Poor nutritional status and inflammation: Serum albumin: Relationship to inflammation and nutrition. *Semin Dial* 17: 432–437. https://doi.org/10.1111/j.0894-0959.2004.17603.x

AIMS Press