Interpretable Machine Learning Model Based on SOFA Score for ICU Sepsis Mortality Prediction with Multicenter Validation

Camilo Santos, Maria A. Bravo, and Carlos A. Fajardo

***Abstract*—The Sequential Organ Failure Assessment (SOFA) score is a widely employed scoring system in clinical practice for predicting mortality in patients with sepsis. The integration of machine learning techniques into clinical scoring systems has enhanced predictive performance; however, many of these models function as "black boxes" offering limited interpretability regarding the contribution of individual clinical variables to the final prediction. This study aims to develop an interpretable machine learning model based on the SOFA score, leveraging its most relevant variables, to predict mortality in Intensive Care Unit (ICU) patients with sepsis using a multicenter validation. The model was trained on data from 15,100 ICU patients in the MIMIC-IV v3.0 dataset and externally validated on 8,201 patients from the eICU v2.0 dataset. The application of an Odds Ratio analysis enabled the identification of the SOFA components demonstrating the strongest association with mortality. This approach facilitated the reduction of variables while enhancing the performance of the model. The interpretability of the model was further addressed by employing SHapley Additive exPlanations (SHAP) values to elucidate the contribution of each variable to the model's predictions. The resulting model outperformed the conventional SOFA score, with improved efficiency and transparency. This interpretable machine learning model, which is based on a SOFA variant, has the potential to support the earlier and more precise intervention required in the clinical management of sepsis ICU patients.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10087

***Index Terms*—Sepsis patients, ICU, machine learning, mortality prediction, multicenter validation, interpretable models.**

I. Introduction

THE Sequential Organ Failure Assessment (SOFA) score is a standardized tool developed to sequentially quantify the degree of organ dysfunction in critically ill patients [1]. Originating from a 1994 consensus conference of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine and first published in 1996 [2]. SOFA evaluates six organ systems: respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems. Each organ system is assigned a score ranging from 0, indicating no dysfunction, to 4, indicating severe dysfunction. Higher total scores are strongly correlated with increased in-hospital mortality [3]. Initially conceived for research purposes, SOFA has since become an integral part of daily intensive care practice, where it guides the monitoring of patient trajectories and estimates the probability of death during Intensive Care Unit (ICU) stay [4].

In the context of sepsis, a condition defined by the Sepsis-3 consensus as life-threatening organ dysfunction caused by a dysregulated host response to infection, the SOFA score is now considered the gold standard for risk stratification and early therapeutic decision-making [1]. Globally, sepsis is a massive public health challenge: in 2017 there were an estimated 48.9 million incident cases and 11.0 million sepsis-related deaths, accounting for roughly 20% of all fatalities worldwide [5].

The fusion of machine learning (ML) with traditional scoring systems like SOFA has yielded marked improvements in predictive metrics [6], but these models often function as "black boxes". Algorithms such as deep neural networks or tree-based ensembles deliver high performance but obscure the individual contribution of each clinical variable, impeding clinician trust and adoption [7]. To overcome this barrier, Explainable AI (XAI) techniques, especially SHapley Additive exPlanations (SHAP), which applies cooperative game theory to allocate the effect of each feature on the model's output, have been developed to illuminate model reasoning [8].

The fusion of machine learning (ML) with traditional scoring systems like SOFA has yielded marked improvements in predictive metrics [6], but these models often function as "black boxes", since their internal decision-making processes are complex and not easily interpretable by humans. Algorithms such as deep neural networks or tree-based ensembles deliver high performance but obscure the individual contribution of each clinical variable, impeding clinician trust and adoption [7]. To overcome this barrier, Explainable AI (XAI) techniques, especially SHapley Additive exPlanations (SHAP), which applies cooperative game theory to allocate the effect of each feature on the model's output, have been developed to illuminate model reasoning [8].

For ML-based risk tools to be reliable in real-world settings, they must undergo rigorous external multicenter validation. Such validation confirms that a model maintains its predictive power across diverse patient populations and clinical environments. Unfortunately, only a minority of published models achieve this level of review, limiting their generalizability and the confidence that healthcare professionals can place in them.

This study constitutes a progression of our previous re-

The associate editor coordinating the review of this manuscript and approving it for publication was Anabel Martin (*Corresponding author: Carlos Fajardo*).

C. Santos, M. A. Bravo, and Carlos Fajardo are with the Department of Electrical, Electronics and Telecommunications, Universidad Industrial de Santander, Bucaramanga, Colombia (e-mails: camilo2238323@correo.uis.edu.co, maria2248087@correo.uis.edu.co, and cafajar@uis.edu.co).

search, entitled "SOFA+: Applying Machine Learning to Enhance the Prognostic Power of the SOFA Score for Predicting ICU Sepsis Mortality" [9]. The present study addresses the limitations of previous ML-augmented SOFA models. SOFA+ employed the original SOFA variables to enhance predictive metrics; however, it did not systematically select the most relevant features, access the interpretability of the model, or employ a multicenter validation. In the demanding ICU environment, reducing the number of input variables has been shown to accelerate computation, streamline data collection, reduce the incidence of missing data, and foster greater acceptance among healthcare professionals [10].

This study aims to develop an interpretable machine learning model based on the SOFA score, leveraging its most relevant variables, to predict mortality in ICU patients with sepsis using a multicenter validation. To identify the most relevant variables, the Odds Ratio (OR) was calculated to assess the association of each variable with mortality in patients with sepsis. To enhance model interpretability, a SHAP analysis was also performed to enable visualization of the individual relationship of each variable with the outcome.

The document begins with a review of the existing literature in Section II. In Section III, we describe the dataset used and the process for selecting the target population based on specific inclusion criteria. This is followed by a detailed explanation of the methodology used to develop a reduced and interpretable model in Section IV. The results are then presented in Section V, highlighting the model's predictive performance in various ICU settings. Finally, in Section VI, the discussion addresses the implications of the findings and outlines the main conclusions in Section VII.

II. Literature Review

Sepsis is one of the leading causes of mortality in the ICU [5], generating a growing interest in the development of predictive models capable of identifying patients with sepsis at higher risk of death in these clinical settings. Traditionally, risk identification has been carried out using scoring systems such as SOFA, SAPS-II, and APACHE [11], which assign a score based on patient severity by evaluating various clinical variables that reflect the functionality and vitality of different organ systems.

Several studies have examined the association between specific factors and mortality in patients with certain conditions in particular contexts using statistical techniques such as the OR and the Hazard Ratio [12]–[17]. These techniques quantify the strength of the association between predictors and outcomes, either by comparing probabilities between groups or by assessing the risk over time. Such statistical methods typically serve as an initial step to identify relevant variables through univariate analyses.

On the other hand, efforts have been made to develop clinical tools aimed at predicting in-hospital mortality, such as the work by Ball I.M et al. [18], in which the authors designed and validated a clinical prediction tool for hospital mortality in critically ill elderly ICU patients, based on OR analysis.

Advances in artificial intelligence and ML techniques have led to the implementation of predictive models that outperform traditional scoring systems in terms of predictive accuracy [19].

Numerous studies have implemented ML models such as Extreme Gradient Boosting (XGBoost), Random Forest, and logistic regression using data from databases such as MIMIC-III, MIMIC-IV, and eICU [20]–[22]. These ML models have demonstrated Area Under the Curve (AUC) values ranging from 0.75 to 0.94, indicating a robust ability to predict sepsis mortality [23]–[28].

For ML-based prognostic tools to be integrated into routine ICU practice, both model interpretability and rigorous multicenter validation are essential [29], [30]. In critical care settings, "black-box" algorithms are of limited use unless they can transparently reveal how patient variables influence risk estimates. Techniques such as SHAP [31] and Local Interpretable Model-Agnostic Explanations (LIME) [32] decompose complex models by attributing risk contributions to clinical variables such as age, Glasgow Coma Scale score, blood urea nitrogen level, and respiratory rate. This clarifies which variables most strongly affect mortality predictions [19], [21], [25], [27]. By illuminating the influence of each feature, clinicians can gain actionable insights into the pathophysiology underlying predicted outcomes and make more informed decisions.

It is equally critical to demonstrate that a prognostic model maintains high performance across diverse ICU populations and care environments. Several studies have externally validated sepsis-mortality models using public and private datasets, demonstrating robust discrimination and calibration between institutions [21], [27], [33].

However, most existing studies are constrained by their reliance on private or single-centre datasets and by the absence of rigorous, variable association analyses with sepsis mortality [29], [30]. To overcome these limitations, we performed a multicenter validation of the SOFA score variables most strongly related to mortality in ICU sepsis patients, quantifying the prognostic contribution of each variable and evaluating its generalizability and clinical impact on the overall risk estimate.

III. Data and Inclusion Criteria

The data were retrospectively obtained from two extensive public databases that contain clinical information of ICU patients. First, we utilized the MIMIC-IV database (version 3.0), which comprises more than 300,000 hospital admissions recorded from 2008 to 2019, for model development [34]. Subsequently, independent testing was performed on the multicenter eICU Collaborative Research database (version 2.0), encompassing 200,859 ICU encounters from 2014 to 2015 [35].

The inclusion selection process involved adult patients admitted to the ICU for at least 24 hours, considering only their first ICU admission. The patients were then selected based on a diagnosis of sepsis according to the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) [36]. Only those with complete demographic records were included, yielding 15,100 sepsis patients from MIMIC-IV and

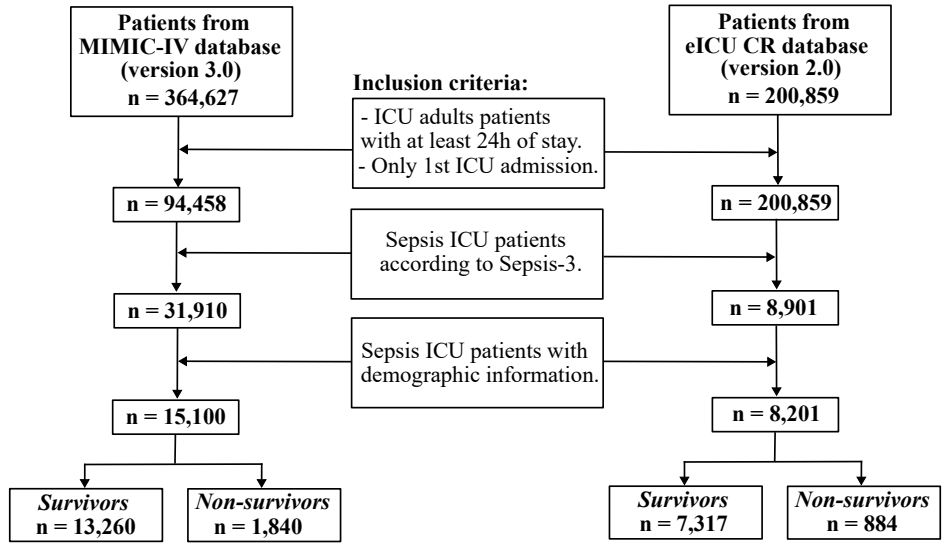

Fig. 1. Inclusion criteria applied to MIMIC-IV and eICU databases for patient selection process.

8,201 from the eICU Collaborative Research database. The patient selection process is detailed in Fig. 1.

After the selection process, we identified the clinical variables that comprise the traditional SOFA score, which total eleven variables. For clinical variables with less than 20% missing data, we applied miceforest [37], a multiple data imputation algorithm based on LightGBM. Miceforest iteratively models each variable with missing data according to the other variables through gradient-boosted trees, capturing complex non-linear relationships and thus estimating the missing data values for each variable.

After the imputation process, we computed the maximum and minimum values for the laboratory tests. Similarly, for the variables corresponding to vital signs, we calculated their maximum, minimum, and mean values. Table I summarizes the twenty-one clinical variables included in our study, which are based on the variables used to develop the widely used SOFA score.

TABLE I
LIST OF TWENTY-ONE CLINICAL VARIABLES EXTRACTED FROM THE MIMIC-IV AND EICU DATABASES

Clinical variables
Mean respiration rate, Min respiration rate, Max respiration rate, Mean mean blood pressure, Min mean blood pressure, Max mean blood pressure, Urine output, Min platelets, Max platelets, Min bilirubin, Max bilirubin, Min creatinine, Max creatinine, Dobutamine, Dopamine, Epinephrine, Norepinephrine, Min Glasgow Coma Scale, Verbal Glasgow Coma Scale, Motor Glasgow Coma Scale, Eyes Glasgow Coma Scale

IV. METHODS

In Section IV-A, we explore a comparative analysis of ML models for mortality prediction in ICU sepsis patients, with the aim of defining the best-performing model, which will seek to reduce the number of variables and perform an interpretability analysis.

To achieve this, an OR analysis is performed in Section IV-B to determine which variables are more strongly associated with mortality in sepsis patients. Finally, Section IV-C explores a SHAP analysis to understand the relationship between each variable and the prediction of mortality in sepsis patients.

A. Comparative Analysis of Machine Learning Models for Mortality Prediction in ICU Sepsis Patients

A total of ten ML models were evaluated to identify the optimal classifier for predicting ICU mortality in adult patients with sepsis, using the AUC as the performance metric. The models included Logistic Regression, Naive Bayes, Support Vector Machines, Decision Trees, Random Forest, Multi-Layer Perceptron, Gradient Boosting Machines, XGBoost, LightGBM, and CatBoost.

For model training, we applied five-fold cross-validation on the MIMIC-IV dataset, recording the AUC for each fold. For external validation, we employed a hold-out approach using the eICU Collaborative Research dataset to enhance the predictive performance of the models evaluated. We selected the model that achieved the highest AUC as the best-performing model.

Fig. 2.a) shows the predictive performance for training in terms of AUC for each model evaluated on the MIMIC-IV dataset, while Fig. 2.b) illustrates the AUC performance for validation of these ML models evaluated on the eICU dataset. Note that some tree-based models, including Random Forest, LightGBM, Gradient Boosting and CatBoost, consistently outperformed other classifiers. Gradient Boosting and CatBoost achieved the highest AUC, with a value of 0.75 (95% CI 0.75 - 0.76) on the eICU dataset. In contrast, the comparatively lower performance of Decision Trees and Support Vector Machines indicates their limited suitability for this prediction task.

Given that Gradient Boosting and CatBoost achieved the same AUC result, we performed Bayesian hyperparameter optimization on both ML models. To enhance transparency and reproducibility, Appendix A provides the details of the

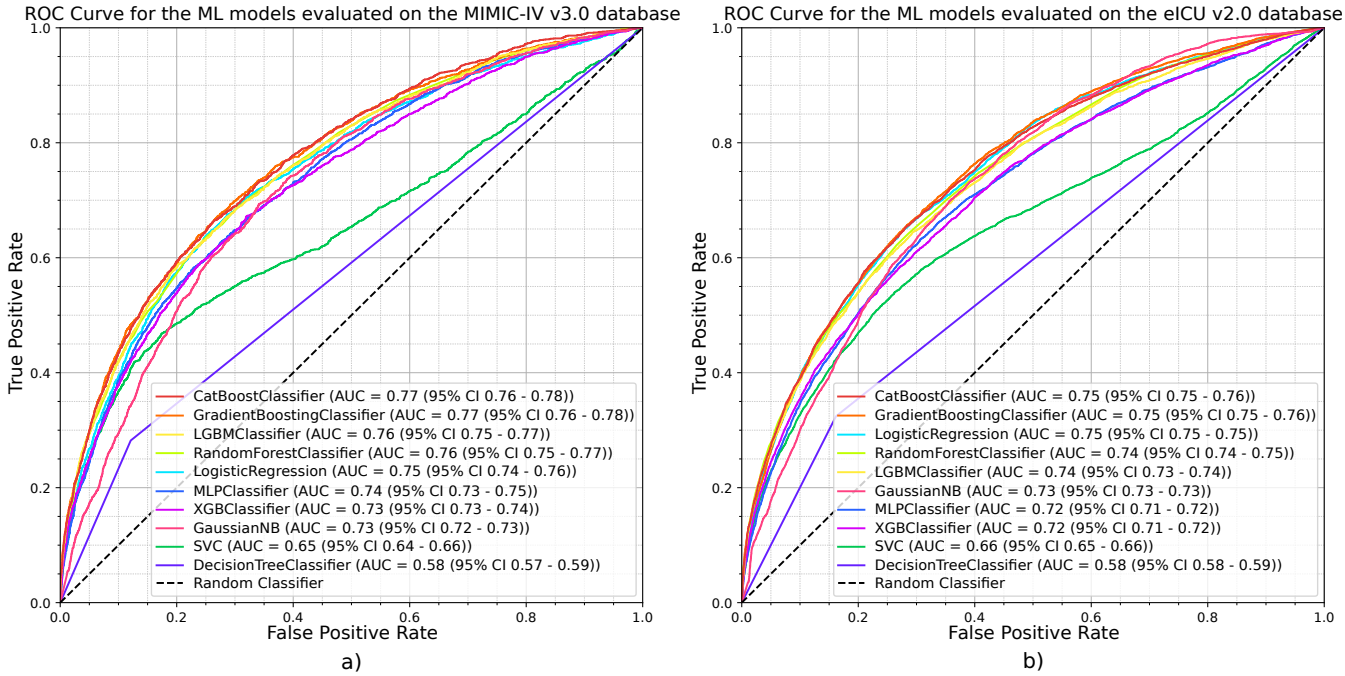

Fig. 2. a) Predictive performance, measured by AUC, of the ten evaluated ML models using five-fold cross-validation on the MIMIC-IV dataset for ICU mortality prediction in sepsis patients. b) Predictive performance, measured by AUC, of the ten ML models evaluated on eICU dataset for ICU mortality prediction in sepsis patients.

TABLE II

PERFORMANCE RESULTS OF THE GRADIENT BOOSTING AND CATBOOST MODEL WITH THE THRESHOLD THAT ACHIEVES THE MAXIMUM F1-SCORE. NOTE: "SPEC": SPECIFICITY, "SENS": SENSITIVITY AND, "F1": F1-SCORE

Models	**MIMIC-IV**				**eICU**			
	AUC (95% CI)	**Spec.**	**Sens.**	**F1**	**AUC (95% CI)**	**Spec.**	**Sens.**	**F1**
CatBoost	0.7783 (0.7717 - 0.7849)	89%	47%	42%	0.7641 (0.7600- 0.7682)	84%	52%	37%
Gradient Boosting	0.7703 (0.7636 - 0.7770)	89%	45%	41%	0.7550 (0.7509 - 0.7592)	84%	50%	36%

Bayesian hyperparameter optimization process. Then we identified the threshold that maximized the F1-score for both models, in order to select the model with the higher sensitivity, which indicates a greater ability to correctly identify patients at risk of ICU mortality. Table II reports the results of AUC, specificity, sensitivity and F1-score obtained for both models using the threshold that achieves the maximum F1-score, demonstrating that the CatBoost model obtained a higher performance.

B. Association of Clinical Variables with ICU Sepsis Mortality

To determine the association between each of the clinical variables included in the SOFA score development and mortality in ICU patients with sepsis, we conducted separate univariable analyses based on logistic regression for the eleven variables. By modeling each predictor, we were able to directly estimate the influence of incremental changes (or the presence or absence of vasopressor use) on the odds of death during a stay in the ICU.

For each analysis, we estimated the odds ratio (OR) to assess the change in mortality risk associated with a one-unit increase (or categorical exposure) of the variable, accompanied by a 95% confidence interval. Variables were then ranked according to the magnitude of their ORs to identify which physiological factors were most strongly associated with mortality in ICU patients with sepsis. These findings highlight potential early warning triggers for clinical intervention.

Subsequently, these results will serve as the basis for developing a reduced version of the model with predictive performance comparable to or superior to the traditional SOFA score, which is used in clinical settings to assess the severity of sepsis in patients.

Based on the OR analysis, different combinations will be proposed in the training of the CatBoost model, progressively omitting some clinical variables. The exclusion of variables will be based on their OR values, as lower values indicate a lower association with mortality.

In the first combination, the CatBoost model will be trained with all clinical variables of the SOFA score except the one with the lowest OR result. Then, successive combinations will be iterated by excluding more additional variables depending on the resulting OR values. This process is repeated until the combinations yield a value equal to or lower than the traditional SOFA score.

This approach will allow us to assess the difference in performance between different proposed combinations and the traditional SOFA score. This will ensure consistency of results

and explore the feasibility of an optimized version of the SOFA score.

C. Interpretability of the Predictive Model

It is crucial to identify the clinical variables influencing the model's predictions, their specific contributions to the outcome and their direction of effect. To address the challenge of interpreting these models and ensuring that our sepsis mortality predictor was transparent and clinically viable, we use one of the most widely recognized approaches to model interpretability in ML: SHapley Additive exPlanations (SHAP) [31].

SHAP quantifies how each input variable contributes, positively or negatively, to individual risk estimates, facilitating the identification of the most influential variables in the model and enhancing trust in these predictive tools.

SHAP provides different types of visualizations depending on the analysis focus, whether at the individual level (per patient) or at the global level (across the entire cohort). One of the most informative visualizations for global analysis is the beeswarm plot, which combines aggregated interpretability with individual-level detail. This plot shows SHAP values for each patient, colored according to the actual value of each variable. This allows variability, impact direction, and potential interactions among variables to be visualized simultaneously. These properties make it a suitable tool for our study, as it enables us to explore and highlight the relationships between the values of clinical variables and their influence on the mortality in ICU patients with sepsis.

By integrating SHAP we provide a transparent and trustworthy model that explains complex ML outputs and facilitates bedside decision-making.

V. Results

A. Validating our Proposed Model through an OR Analysis

Table III shows the association of each individual variable comprising the traditional SOFA score with mortality in ICU patients with sepsis, ranked in descending order based on their OR values.

TABLE III
Odds Ratio Analysis for Individual Clinical Variables of SOFA Score

SOFA variables	OR	95% CI
Urine output	1.6989	(1.6534 - 1.7458)
Creatinine	1.5268	(1.4813 - 1.5736)
Mean Blood Pressure	1.5072	(1.4716 - 1.5436)
Platelets	1.4972	(1.4609 - 1.5344)
Respiration rate	1.4277	(1.3925 - 1.4638)
Bilirubin	1.3292	(1.2950 - 1.3644)
Glasgow Coma Scale	1.3077	(1.2786 - 1.3374)
Norepinephrine	1.3054	(1.2699 - 1.3419)
Epinephrine	1.2838	(1.2588 - 1.3092)
Dobutamine	1.1216	(1.0967 - 1.1471)
Dopamine	1.0314	(1.0027 - 1.0608)

Based on the results of the OR analysis, note that urine output is the clinical variable most strongly associated with mortality in ICU sepsis patients (OR = 1.6989, 95% CI: 1.6534–1.7458). This indicates that the odds of mortality increase by around 1.7 times in association with changes in urine output, underscoring its critical role as a predictor in this context.

In contrast, the clinical variables that are least associated with the outcome belong to the vasopressor use category. These are norepinephrine, epinephrine, dobutamine and dopamine. Of these, norepinephrine shows the strongest association, while dopamine shows the weakest.

Based on this finding, we proposed different combinations to assess the impact of sequentially excluding variables, beginning with the one demonstrating the weakest association with mortality in ICU patients with sepsis. The first combination, referred to as Model 1, involved training our CatBoost model with all variables except dopamine. Next, in Model 2, we excluded both dopamine and dobutamine variables. Subsequently, according to our results in Table III, Model 3 omitted the next variable, epinephrine. In this model, CatBoost was trained without dopamine, dobutamine, and epinephrine variables. Following the same approach, Model 4 excluded norepinephrine in addition to the previous three variables, thus removing the entire vasopressor category.

Further analyzing the OR results, the next variable was the Glasgow Coma Scale (OR = 1.3077, 95% CI: 1.2786–1.3374), which is a neurological scale used to assess a patient's level of consciousness. We trained our CatBoost model excluding this variable along with the four previously removed ones, referring to this combination as Model 5. The next combination excluded bilirubin variable (OR = 1.3292, 95% CI: 1.2950–1.3644), resulting in Model 6. Finally, Model 7 involved training the CatBoost model by excluding the respiration rate variable (OR = 1.4277, 95% CI: 1.3925–1.4638) in addition to the variables removed in previous models.

Fig. 3 illustrates how the different models were constructed, following the previously described methodology, where variables were progressively excluded based on the OR analysis results.

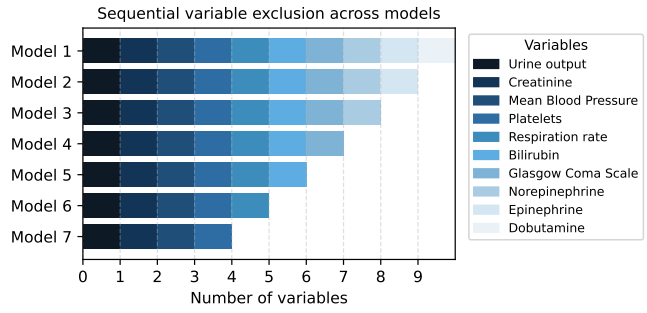

Fig. 3. Different model combinations constructed by sequentially excluding variables based on OR analysis.

The results obtained from each of these experiments are summarized in Table IV, which compares the performance metrics of each proposed combination, determined by OR analysis, with the traditional SOFA score.

TABLE IV
PERFORMANCE RESULTS OF THE DIFFERENT COMBINATIONS WITH CATBOOST MODEL USING A REDUCED NUMBER OR CLINICAL SOFA VARIABLES. MODEL 1 EXCLUDES DOPAMINE VARIABLE. MODEL 2 EXCLUDES DOPAMINE AND DOBUTAMINE VARIABLES. MODEL 3 EXCLUDES DOPAMINE, DOBUTAMINE AND EPINEPHRINE VARIABLES. MODEL 4 EXCLUDES DOPAMINE, DOBUTAMINE, EPINEPHRINE AND NOREPINEPHRINE VARIABLES, WHICH CONSISTS IN VASOPRESSORS USE CATEGORY. MODEL 5 EXCLUDES VASOPRESSORS AND GLASGOW COMA SCALE (GCS) VARIABLES. MODEL 6 EXCLUDES VASOPRESSORS, GCS AND BILIRUBIN VARIABLES. MODEL 7 EXCLUDES VASOPRESSORS, GCS, BILIRUBIN AND RESPIRATION RATE VARIABLES. NOTE: "SPEC": SPECIFICITY, "SENS": SENSITIVITY AND, "F1": F1-SCORE

Models	MIMIC-IV				eICU			
	AUC (95% CI)	Spec.	Sens.	F1	AUC (95% CI)	Spec.	Sens.	F1
SOFA score	0.6904 (0.6830 - 0.6978)	82%	44%	32%	0.7228 (0.7131 - 0.7325)	90%	36%	33%
Model 1	0.7792 (0.7725 - 0.7858)	89%	47%	41%	0.7659 (0.7618 - 0.7700)	84%	52%	37%
Model 2	0.7788 (0.7722 - 0.7854)	89%	47%	41%	0.7642 (0.7600 - 0.7683)	84%	52%	37%
Model 3	0.7792 (0.7726 - 0.7858)	89%	48%	42%	0.7627 (0.7586 - 0.7668)	84%	52%	37%
Model 4	**0.7753 (0.7686 - 0.7819)**	**86%**	**52%**	**41%**	**0.7647 (0.7606 - 0.7688)**	**79%**	**58%**	**35%**
Model 5	0.7342 (0.7272 - 0.7413)	88%	43%	37%	0.7528 (0.7486 - 0.7569)	72%	66%	33%
Model 6	0.7295 (0.7224 - 0.7366)	89%	40%	37%	0.7495 (0.7453 - 0.7537)	73%	64%	33%
Model 7	0.7052 (0.6979 - 0.7125)	88%	39%	35%	0.7209 (0.7166 - 0.7253)	76%	56%	32%

According to the results presented in Table IV, it can be observed that all the proposed combinations outperform the traditional SOFA score, with the exception of Model 7, which achieves an AUC of 0.7209 (95% CI: 0.7166 - 0.7253), slightly lower than the AUC achieved by the SOFA score [0.7228 (95% CI: 0.7131 - 0.7325)] on the eICU dataset.

We found that removing only the dopamine variable (Model 1) led to a significant improvement in performance compared to the traditional SOFA score baseline. The AUC increased from 0.6904 (95% CI: 0.6830 - 0.6978) to 0.7792 (95% CI: 0.7725 - 0.7858) in MIMIC-IV dataset and from 0.7228 (95% CI: 0.7131 - 0.7325) to 0.7659 (95% CI: 0.7618 - 0.7700) in eICU dataset.

Interestingly, Model 2, Model 3 and Model 4, progressively exclude additional vasopressor-related variables (dobutamine, epinephrine, and norepinephrine), exhibit comparable or slightly improved predictive performance compared to Model 1. This suggests that despite their clinical relevance, vasopressor variables may not contribute substantially to mortality prediction in this context.

In contrast, Model 5, Model 6 and Model 7, which exclude variables beyond the vasopressor category as well as other clinical variables, demonstrate a progressive decline in model performance. The exclusion of clinical variables such as the Glasgow Coma Scale, bilirubin and respiration rate leads to decreased AUC, sensitivity and F1-score. Notably, Model 7, which excludes a larger number of variables, in particular, shows lower performance than the traditional SOFA score. Nevertheless, it could represent a viable alternative in ICU settings with limited access to clinical data.

These findings underscore the value of a data-driven variable selection strategy guided by OR analysis. Furthermore, they emphasize the trade-off between model simplicity and predictive power. Removing variables with low ORs can enhance model efficiency and interpretability without compromising and in some cases even improving predictive performance.

B. Model Explainability: SHAP Results

According to the results presented in Table IV, Model 4, which excludes the four clinical variables related to vasopressor use, was identified as the best-performing model, achieving the highest sensitivity in MIMIC-IV dataset among all evaluated combinations.

To analyze the interpretability of this model, we used a SHAP-based beeswarm plot to examine the individual contribution of each clinical variable to the prediction of mortality in ICU patients with sepsis. Fig. 4 shows the beeswarm plot with the SHAP values for each clinical variable in Model 4, ranked according to their importance in the model.

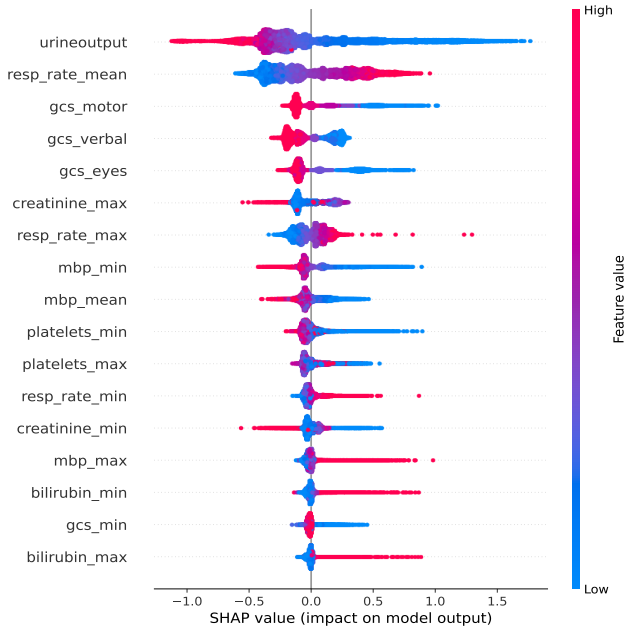

Fig. 4. Beeswarm plot for SHAP analysis of the input clinical variables in Model 4.

According to the SHAP analysis, urine output was the most influential variable in the mortality prediction model, followed

by the mean respiration rate and the three components of the Glasgow Coma Scale (motor, verbal, and eye response). These findings are consistent with the results of the OR analysis presented in Table III, which showed that urine output was most strongly associated with mortality.

Specifically, high levels of urine output contributed significantly to a lower predicted risk of mortality, while low levels increased the estimated risk, according to the model. In contrast, higher values of the mean respiration rate contributed to an increased predicted risk of mortality, whereas lower values generally decreased the predicted risk, indicating an inverse relationship compared to urine output.

Similarly, neurological status abnormalities (as captured by the GCS components) were important contributors to the model's predictions. Additionally, laboratory parameters such as creatinine, platelet count, and bilirubin showed moderate importance in the predictive model.

VI. Discussion

This study presents a reduced and interpretable ML model for predicting mortality in ICU patients with sepsis, which can help to improve patient care and outcomes. The model is based on the clinical variables that comprise the widely used SOFA score and has been validated across diverse ICU settings, highlighting its robustness and potential for use in clinical practice.

Among the ten ML models evaluated, CatBoost demonstrated the best predictive performance. Its ability to handle categorical variables and capture non-linear relationships likely contributed to its enhanced accuracy in this clinical context.

From the association analysis performed on individual variables through an OR analysis, urine output emerged as the most strongly associated predictor of mortality, with the highest value (OR = 1.6989, 95% CI: 1.6534–1.7458). This finding aligns with clinical expectations, as urine output is a well-established marker in the diagnosis and progression of sepsis. This strong association is consistent with the literature, especially considering that urinary tract infections can progress to sepsis if not treated promptly [38].

In contrast, vasopressor-related variables exhibited the weakest associations. Of these, norepinephrine had the highest OR, while dopamine had the lowest. These results are coherent with current clinical guidelines, which recommend norepinephrine as the primary vasopressor for managing septic shock due to its proven efficacy and favorable safety profile [39].

Based on this analysis, we explored different model combinations, progressively excluding the variables with the lowest associations, starting with the dopamine variable and assessing the impact on model performance. Each subsequent model removed the variables from the previous combination along with the next least associated variable. This strategy revealed that our ML approach consistently outperformed the traditional SOFA score, even when certain variables were omitted, as can be seen in all combinations presented in Table IV. Based on these results, the only exception was Model 7, in which the SOFA score slightly outperformed our proposed model in the eICU dataset. However, the overall approach enabled the development of a compact model that retained only the most relevant predictors while maintaining high predictive performance for estimating the ICU mortality risk.

One notable result came from Model 1, which excluded only the dopamine variable. Its performance suggests that dopamine is not only a weak predictor for mortality, but its inclusion may introduce noise or redundancy that affects model performance. As more vasopressor-related variables were removed, performance metrics — particularly sensitivity — improved in the MIMIC-IV dataset and remained stable in the eICU dataset. Model 4, which excluded all four vasopressor-related variables, achieved the best balance between these performance metrics.

It is important to highlight the methodological differences between SHAP-based interpretability analysis and odds ratio analysis. In the SHAP analysis, multiple statistical descriptors (e.g., minimum, maximum, mean) of each variable were considered independently, as detailed in Table I. In contrast, the OR analysis grouped these statistical measures under a single clinical variable. Therefore, a direct comparison between the two analyses is not feasible. Nevertheless, both methods consistently identified urine output as the most important and most strongly associated variable. This agreement is likely due to the fact that urine output is represented as a single raw variable rather than through multiple statistical descriptors.

A notable strength of our study is the external validation achieved through implementation in a multicenter dataset. This enhances the generalizability and robustness of our findings, representing a significant improvement over previous studies that frequently relied on a single dataset for both model development and validation.

VII. Conclusions

In this study, a reduced and interpretable variant of the SOFA score is presented, with machine learning being leveraged to predict ICU mortality in sepsis patients. The application of Odds-Ratio–based feature selection to the original SOFA components resulted in a reduction of the model's dimensionality without any compromise to performance. Our model outperformed the conventional SOFA score in both the MIMIC-IV and eICU cohorts.

The integration of SHAP values is crucial to providing clinicians with clear, example-level explanations of how each organ-system variable influences individual risk estimates. This addresses the 'black-box' barrier that often limits the uptake of machine learning tools in high-stakes ICU settings. The model's reduced set of inputs streamlines data collection, minimizes missingness, and facilitates real-time deployment within electronic health record–driven alert systems.

The interpretable reduced machine learning-based SOFA model under consideration is a robust and generalizable tool for the early risk stratification of ICU patients suffering from sepsis. It has the potential to support data-driven clinical decisions in a timely manner, which would ultimately improve the survival rate of this population.

To acknowledge the limitations of our study, it is important to note that the evaluation was conducted using publicly available datasets (MIMIC-IV and eICU). While these databases are widely used and provide high-quality data, their characteristics may not fully capture the heterogeneity of clinical practices across different healthcare settings, particularly in hospitals with limited resources or distinct care protocols. Therefore, the generalizability of our findings to such environments may be constrained. Future work should consider validating the proposed models in diverse limited clinical contexts to ensure broader applicability and robustness.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the Universidad Industrial de Santander (UIS) for the institutional support provided during the development of this work, as well as for the access to academic, computational, and physical resources necessary to conduct the research. This project was made possible thanks to the academic environment that fosters research and comprehensive student training.

APPENDIX
HYPERPARAMETER OPTIMIZATION DETAILS

Table V summarizes the search ranges considered during the hyperparameter tuning process and reports the final optimal values. A Bayesian optimization strategy was employed to efficiently navigate the hyperparameter space, yielding the configuration that maximized predictive performance.

TABLE V
SEARCH RANGES AND OPTIMAL VALUES FOR THE BAYESIAN HYPERPARAMETER OPTIMIZATION PROCESS APPLIED IN THIS STUDY

Hyperparameter	Search range	Optimal value
iterations	(100, 500)	343
learning_rate	(0.01, 0.3)	0.0313
depth	(3, 10)	6
l2_leaf_reg	(1, 10)	5.1552
random_strength	(0.1, 5)	1.5877

REFERENCES

[1] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, Feb. 2016. DOI: 10.1001/jama.2016.0287

[2] R. Moreno *et al.*, "The Sequential Organ Failure Assessment (SOFA) Score: has the time come for an update?," *Critical Care*, vol. 27, no. 1, p. 15, Jan. 2023. DOI: 10.1186/s13054-022-04290-9

[3] A. M. Khan and S. M. Aslam, "Comparison of qSOFA Score, SIRS Criteria, and SOFA Score as predictors of mortality in patients with sepsis," *Ghana Medical Journal*, vol. 56, no. 3, pp. 191–197, Sep. 2022. DOI: 10.4314/gmj.v56i3.9

[4] E. P. Raith *et al.*, "Prognostic Accuracy of the SOFA Score, SIRS Criteria, and qSOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit," *JAMA*, vol. 317, no. 3, pp. 290–300, Jan. 2017. DOI: 10.1001/jama.2016.20328

[5] K. E. Rudd *et al.*, "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study," *Lancet*, vol. 395, no. 10219, pp. 200–211, Jan. 2020. DOI: 10.1016/S0140-6736(19)32989-7

[6] X. Pan, J. Xie, L. Zhang *et al.*, "Evaluate prognostic accuracy of SOFA component score for mortality among adults with sepsis by machine learning method," *BMC Infect Dis*, vol. 23, p. 76, 2023. DOI: 10.1186/s12879-023-08045-x

[7] T. A. A. Abdullah, M. S. M. Zahid, and W. Ali, "A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions," *Symmetry*, vol. 13, no. 12, article 2439, 2021. DOI: 10.3390/sym13122439

[8] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020. DOI: 10.3390/e23010018

[9] C. Santos, M. A. Bravo, and C. A. Fajardo, "SOFA+: Applying Machine Learning to Enhance the Prognostic Power of the SOFA Score for Predicting ICU Sepsis Mortality," in *Proc. 2025 XXV Symp. Image, Signal Processing, and Artificial Vision (STSIVA)*, Bucaramanga, Colombia, 2025, pp. 1–5. DOI: 10.1109/STSIVA66383.2025.11156628.

[10] A. E. W. Johnson *et al.*, "Machine Learning and Decision Support in Critical Care," *Proc. IEEE*, vol. 104, no. 2, pp. 444–466, 2016. DOI: 10.1109/JPROC.2015.2501978

[11] J.-L. Vincent *et al.*, "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, 1996. DOI: 10.1007/BF01709751

[12] J. Yang, L. Li, and S. Fu, "Association analysis of sepsis progression to sepsis-induced coagulopathy: a study based on the MIMIC-IV database," *BMC Infectious Diseases*, vol. 25, no. 1, p. 573, 2025. DOI: 10.1186/s12879-025-10972-w

[13] M. S. Ki *et al.*, "Association Between Plasma Granzyme B Levels, Organ Failure, and 28-Day Mortality Prediction in Patients with Sepsis," *Journal of Clinical Medicine*, vol. 14, no. 5, p. 1461, 2025. DOI: 10.3390/jcm14051461

[14] G. L. Sacha *et al.*, "Association of Catecholamine Dose, Lactate, and Shock Duration at Vasopressin Initiation With Mortality in Patients With Septic Shock," *Critical Care Medicine*, vol. 50, no. 4, pp. 614–623, 2022. DOI: 10.1097/CCM.0000000000005317

[15] I. D. Saragih, I. S. Saragih, S. O. Batubara, and C.-J. Lin, "Dementia as a mortality predictor among older adults with COVID-19: A systematic review and meta-analysis of observational study," *Geriatric Nursing*, vol. 42, no. 5, pp. 1230–1239, 2021. DOI: 10.1016/j.gerinurse.2021.03.007

[16] H. Estiri, Z. H. Strasser, J. G. Klann, P. Naseri, K. B. Wagholikar, and S. N. Murphy, "Predicting COVID-19 mortality with electronic medical records," *NPJ Digital Medicine*, vol. 4, no. 1, p. 15, 2021, DOI: 10.1038/s41746-021-00383-x.

[17] N. Charoenngam, A. Shirvani, N. Reddy, D. M. Vodopivec, C. M. Apovian, and M. F. Holick, "Association of vitamin D status with hospital morbidity and mortality in adult hospitalized patients with COVID-19," *Endocrine Practice*, vol. 27, no. 4, pp. 271–278, 2021. DOI: 10.1016/j.eprac.2021.02.013

[18] I. M. Ball, S. M. Bagshaw, K. E. A. Burns, D. J. Cook, A. G. Day, P. M. Dodek, D. J. Kutsogiannis, S. Mehta, J. G. Muscedere, H. T. Stelfox, A. F. Turgeon, G. A. Wells, and I. G. Stiell, "A clinical prediction tool for hospital mortality in critically ill elderly patients," *Journal of Critical Care*, vol. 35, pp. 206–212, 2016. DOI: 10.1016/j.jcrc.2016.05.026.

[19] X. Q. Luo, P. Yan, S. B. Duan, Y. X. Kang, Y. H. Deng, Q. Liu, T. Wu, and X. Wu, "Development and validation of machine learning models for real-time mortality prediction in critically ill patients with sepsis-associated acute kidney injury," *Frontiers in Medicine*, vol. 9, 2022. DOI: 10.3389/fmed.2022.853102.

[20] S. Shi, L. Zhang, S. Zhang, J. Shi, D. Hong, S. Wu, X. Pan, and W. Lin, "Developing a rapid screening tool for high-risk ICU patients of sepsis: integrating electronic medical records with machine learning methods for mortality prediction in hospitalized patients—model establishment, internal and external validation, and visualization," *Journal of Translational Medicine*, vol. 23, no. 1, 2025. DOI: 10.1186/s12967-025-06102-4.

[21] G. Zhang, F. Shao, W. Yuan, J. Wu, X. Qi, J. Gao, R. Shao, Z. Tang, and T. Wang, "Predicting sepsis in-hospital mortality with machine learning: a multi-center study using clinical and inflammatory biomarkers," *European Journal of Medical Research*, vol. 29, no. 1, 2024, DOI: 10.1186/s40001-024-01756-0.

[22] J. A. Castillo, C. Santos, and C. A. Fajardo, "Multicenter validation of a machine learning algorithm for mortality prediction at decision-critical timestamps in ICU sepsis patients," in *Proceedings of the 2025 8th International Conference on Robot Systems and Applications (ICRSA 2025)*, Wuhan, China, Sept. 19–21, 2025, pp. 1–5. Article accepted.

[23] M. S. Rahman, K. R. Islam, J. Prithula, J. Kumar, M. Mahmud, M. F. Alam, M. B. I. Reaz, A. Alqahtani, and M. E. H. Chowdhury, "Machine learning-based prognostic model for 30-day mortality prediction in Sepsis-3," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, 2024, DOI: 10.1186/s12911-024-02655-4.

[24] A. Shumilov, Y. Zhu, N. Ashrafi, A. Abdollahi, G. Placencia, K. Alaei, and M. Pishgar, "Data-Driven Machine Learning Approaches for Predicting In-Hospital Sepsis Mortality," *Lecture Notes in Networks and Systems*, vol. 1424 LNNS, pp. 393–409, 2025, DOI:10.1007/978-3-031-92605-1_24.

[25] C. Hu, L. Li, W. Huang, T. Wu, Q. Xu, J. Liu, and B. Hu, "Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study," *Infectious Diseases and Therapy*, vol. 11, no. 3, pp. 1117–1132, 2022, DOI: 10.1007/s40121-022-00628-6.

[26] S. Li, R. Dou, X. Song, K. Y. Lui, J. Xu, Z. Guo, X. Hu, X. Guan, and C. Cai, "Developing an interpretable machine learning model to predict in-hospital mortality in sepsis patients: A retrospective temporal validation study," *Journal of Clinical Medicine*, vol. 12, no. 3, 2023, Art. no. 915. DOI: 10.3390/jcm12030915.

[27] J. Zhuang, H. Huang, S. Jiang, J. Liang, Y. Liu, and X. Yu, "A generalizable and interpretable model for mortality risk stratification of sepsis patients in intensive care unit," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, 2023, Art. no. 2279. DOI: 10.1186/s12911-023-02279-0.

[28] L. Shen, J. Wu, J. Lan, C. Chen, Y. Wang, and Z. Li, "Interpretable machine learning-based prediction of 28-day mortality in ICU patients with sepsis: a multicenter retrospective study," *Frontiers in Cellular and Infection Microbiology*, vol. 14, 2024, Art. no. 1500326. DOI: 10.3389/fcimb.2024.1500326.

[29] L. Mondrejevski, F. Rugolon, I. Miliou, and P. Papapetrou, "MASICU: A multimodal attention-based classifier for sepsis mortality prediction in the ICU," in *Proc. IEEE Symp. Computer-Based Medical Systems (CBMS)*, 2024, pp. 326–331. DOI: 10.1109/CBMS61543.2024.00061.

[30] C. Q. Zhu, M. Tian, L. Semenova, J. Liu, J. Xu, J. Scarpa, and C. Rudin, "Fast and interpretable mortality risk scores for critical care patients," *J. Amer. Med. Informatics Assoc.*, vol. 32, no. 4, pp. 736–747, 2025. DOI: 10.1093/jamia/ocae318.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. DOI: 10.48550/arXiv.1705.07874.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[33] C. Santos, J. X. Ramos Garzón, S. D. Pertuz, and C. A. Fajardo, "Significant clinical factors for mortality prediction in ICU sepsis patients: A machine learning approach," *Smart Health*, vol. 38, p. 100613, 2025. DOI: 10.1016/j.smhl.2025.100613.

[34] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023. DOI: 10.1038/s41597-022-01899-x.

[35] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, no. 1, p. 180178, 2018. DOI: 10.1038/sdata.2018.178.

[36] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J.-L. Vincent, and D. C. Angus, "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, Feb. 2016. DOI: 10.1001/jama.2016.0287.

[37] The miceforest Development Team, *miceforest: Fast, Memory Efficient Imputation with LightGBM*, versión 6.0.3, 2024. [Online]. Available: https://pypi.org/project/miceforest/

[38] A. Porat, B. S. Bhutta, and S. Kesler, "Urosepsis," in *StatPearls [Internet]*, Treasure Island (FL): StatPearls Publishing, 2025. [Updated: Aug. 17, 2023]. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK482344/. PMID: 29493969. Bookshelf ID: NBK482344.

[39] T. Avni, A. Lador, S. Lev, L. Leibovici, M. Paul, and A. Grossman, "Vasopressors for the treatment of septic shock: Systematic review and meta-analysis," *PLOS ONE*, vol. 10, no. 8, p. e0129305, 2015, doi: 10.1371/journal.pone.0129305.

Camilo Santos received his B.Sc. degree in Electronic Engineering from the Universidad Industrial de Santander (UIS), Colombia. He is currently awaiting the award of his M.Sc. degree in Telecommunications Engineering from UIS, where he is also pursuing a Ph.D. in Engineering. He is a member of the Connectivity and Signal Processing (CPS) Research Group at UIS. His research interests include the application of artificial intelligence to intensive care unit environments, with a particular focus on interpretable machine learning models for clinical decision support. During his master's program, he was awarded a full scholarship covering tuition and living expenses.

Maria A. Bravo received her B.Sc. degree in Electronic Engineering from the Universidad Industrial de Santander (UIS), Colombia. She is currently in the final semester of her M.Sc. degree in Electronic Engineering at UIS. She is a member of the Connectivity and Signal Processing (CPS) Research Group at UIS. Her research interests focus on the application of artificial intelligence to pediatric intensive care unit environments, with an emphasis on developing interpretable machine learning models for clinical decision support. During her master's program, she was awarded a full scholarship covering tuition and living expenses.

Carlos A. Fajardo is a faculty professor at Universidad Industrial de Santander (UIS), Colombia. He holds a Ph.D. in Engineering with a focus on High-Performance Computing, an M.Sc. in Electronic Engineering with a specialization in Advanced Digital Design, and a postgraduate certificate in University Teaching, all from UIS. He completed a postdoctoral fellowship at the Center for Brain-Inspired Computing (C-BRIC) at Purdue University, where he specialized in edge AI through hardware-software co-design, and also served as a visiting researcher at Purdue's Integration Lab. His research focuses on artificial intelligence applied to medical problems, with additional expertise in advanced digital systems and hardware-software co-design.