Scientific Reports

https://doi.org/10.1038/s41598-025-34385-6

Article in Press

Hybrid phenotype-guided modeling across algorithm–feature regimes with application to ICU mortality prediction for *Acinetobacter baumannii*

Received: 8 October 2025

Accepted: 29 December 2025

Published online: 03 January 2026

Cite this article as: Eid A., Khayyat R., Alwafa S.A. *et al.* Hybrid phenotype-guided modeling across algorithm–feature regimes with application to ICU mortality prediction for *Acinetobacter baumannii*. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-34385-6

Abdelrahman Eid, Rasha Khayyat, Shadi Abu Alwafa & Hadi Rabee

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Hybrid phenotype-guided modeling across algorithm–feature regimes with application to ICU mortality prediction for *Acinetobacter baumannii*

Abdelrahman Eid, PhD[1,2*], Rasha Khayyat, PhD[3,4*], Shadi Abu Alwafa, MS[2], Hadi Rabee, MD[5]

1 Department of Mathematics, An-Najah National University, Nablus P.O. Box 7, Palestine
2 Data Science Unit, An-Najah Innovation Park, An-Najah National University, Nablus P.O. Box 7, Palestine
3 Department of Biomedical Sciences and Basic Clinical Skills , An-Najah National University, Nablus P.O. Box 7, Palestine
4 Medical Sciences Research Unit, Scientific Centers, An-Najah National University, Nablus P.O. Box 7, Palestine
5 Department of Medicine, An-Najah National University, Nablus P.O. Box 7, Palestine

* Correspondence: Abdelrahman Eid, abed.eid@najah.edu (A.E.)
Rasha Khayyat, rasha.khayyat@najah.edu (R.K.)

Abstract

Acinetobacter baumannii causes severe Intensive Care Unit (ICU) infections with high mortality, yet most prediction tools rely on risk scores or supervised machine learning (ML) and overlook hidden patient subgroups. This study applied a hybrid machine learning framework combining unsupervised clustering and supervised prediction to refine ICU mortality estimation and evaluate whether incorporating phenotype information enhances performance. Patient phenotypes were identified using clustering, and cluster membership was incorporated as an additional predictive feature. ML models were trained with and without cluster membership under two feature settings: full-feature models including all variables and reduced-feature models limited to significant predictors identified within the clusters. The resulting phenotypes were clinically distinct and strongly associated with mortality, demonstrating that data-driven patient grouping can provide complementary prognostic information. Incorporating phenotype membership improved predictive accuracy in a context-dependent manner, varying by feature regime and the learning algorithm. This study introduces a novel framework for phenotype-guided critical care modeling that bridges unsupervised and supervised learning, advancing personalized critical care and supporting global efforts to reduce preventable ICU mortality.

Keywords: Machine learning in critical care, ICU mortality prediction, Phenotype-guided modeling, Clinical risk stratification, *Acinetobacter baumannii* prediction, Unsupervised clustering.

1 Introduction

Acinetobacter baumannii is a major cause of hospital-acquired infections in intensive care units. Its ability to survive in the hospital environment and to rapidly acquire antimicrobial resistance has made it a critical-priority pathogen on the World Health Organization (WHO) list [1,2].

Mortality rates among ICU patients with A. *baumannii* infections are strikingly high, ranging from 10% to more than 70% depending on resistance patterns and timeliness of effective treatment [3,4].

Prognosis is typically assessed using severity scores, such as Sequential Organ Failure Assessment (SOFA) and Acute Physiologic Assessment and Chronic Health Evaluation (APACHE), or supervised ML models that use large sets of clinical variables [5,6]. While these approaches can predict outcomes with moderate accuracy, they generally treat all patients as a single homogeneous group, and ignore heterogeneity in comorbidities, infection sites, and responses to treatment, which may mask important subgroups at higher or lower risk of mortality.

In broader ICU research, unsupervised learning methods such as clustering and latent class analysis have been successfully applied to identify patient phenotypes that differ in outcomes and even in treatment response. Phenotypes of sepsis and septic shock have been linked to 14-day mortality and organ failure trajectories [7,8]. Similarly, subgroups in acute respiratory distress syndrome (ARDS) showed differential mortality and response to therapy [9]. These advances demonstrate that machine-learned phenotypes can provide clinically meaningful stratification. However, no studies to date have applied such methods to ICU patients with A. *baumannii* infection.

ML is increasingly applied in intensive care medicine, where supervised models have shown strong performance in predicting sepsis, organ failure, and short-term mortality [5,10].ICU patients with *A. baumannii* infection are one such group, typically having multiple comorbidities, severe organ dysfunction, and intensive use of invasive support and antibiotics, making them a heterogeneous, high-risk population well suited to ML-based analysis.Yet, most models still assume patient populations are uniform and rarely account for hidden subgroups that may differ in risk or response. Similarly, while clustering and other unsupervised methods have been applied to broader ICU cohorts, little is known about how such phenotype labels interact with supervised algorithms to improve prediction. This represents an important methodological gap: the role of unsupervised phenotypes as features in supervised learning has not been systematically tested across modeling strategies or feature regimes.

To address this gap, we designed a two-stage study with complementary clinical and ML objectives. Clinically, we aimed to uncover phenotypes of ICU patients with A. *baumannii* infection and test their association with mortality. From an ML perspective, we examined whether incorporating phenotype labels into supervised models improves predictive performance, and under what conditions such benefits arise. Methodologically, we treat this as a phenotype-guided modelling framework: a model-agnostic, stepwise pipeline that (i) derives phenotypes using unsupervised clustering on routine ICU variables, (ii) integrates the phenotype label as an optional predictor in different algorithms and feature sets, and (iii) compares performance across these regimes to quantify the incremental value of phenotype information. This study explores when unsupervised phenotypes add value to supervised prediction, and contributes not only to infection-specific critical care but also to the broader question of how clustering-derived features can enhance clinical ML.

Given the high case-fatality of ICU *A. baumannii* infection, we focused on in-hospital mortality as the primary outcome, using predictors available during the index ICU admission to enable early risk stratification and to inform decisions about monitoring, antimicrobial therapy, and organ support.

2 Literature Review

Early attempts to predict outcomes in A. *baumannii* infection relied on conventional severity scores such as SOFA and APACHE, or regression-based analyses of clinical risk factors. These studies identified illness severity, comorbidities, and invasive procedures as predictors of poor prognosis, but their discriminative ability was modest and inconsistent across cohorts [3]. Logistic regression models, including nomograms, were later developed to refine risk prediction, incorporating variables such as infection source, mechanical ventilation, albumin, and comorbidity burden. While these improved calibration, their discriminatory performance remained limited, with Area Under the Curve (AUCs) generally below 0.80 [12].

With the rise of ML, more flexible models were introduced. Xu et al. applied an interpretable gradient boosting framework (XGBoost) to predict fulminant sepsis due to A. *baumannii* bloodstream infection and demonstrated better performance than conventional severity scores [5]. Neuman et al. developed and externally validated a prediction model for hospital-acquired A. *baumannii* using electronic health record data, confirming feasibility but with only modest discrimination [13]. Other contemporary studies combined regression and ML to address carbapenem-resistant A. *baumannii* bloodstream infections [11,4]. These studies illustrate the transition from traditional regression-based models to supervised ML approaches, though most continue to treat patients as a homogeneous group and lack strategies to address underlying heterogeneity.

In parallel, unsupervised methods have been applied more broadly in critical care to uncover latent clinical subgroups. In ARDS, latent class analysis revealed “hyper-inflammatory” and “hypo-inflammatory” phenotypes with distinct mortality and treatment responses [9]. Subsequent studies validated these subphenotypes and showed they can be predicted using routine clinical variables, enabling practical bedside classification [15,16]. Similar progress was made in sepsis, where Seymour et al. identified four phenotypes associated with unique host-response patterns and outcome trajectories [7]. More recent studies extended this approach to sepsis-associated ARDS, where mortality varied dramatically across subgroups [14]. Unsupervised clustering has revealed clinically meaningful phenotypes across sepsis, ARDS, and COVID-19, consistently associated with prognosis and treatment response [17–19].

Supervised ML methods such as Random Forest (RF) and gradient boosting (including XGBoost) are now well established in ICU prognostic modeling, owing to their ability to capture nonlinearities, interactions, and high-dimensional data while providing interpretable feature importance [20–22]. For unsupervised learning, Partitioning Around Medoids (PAM) has been recommended in clinical contexts because it accommodates mixed variable types and is more robust to outliers than k-means [23,24]. Prior successes of these approaches in ICU research therefore provide strong justification for their use in our study: RF and XGBoost for mortality prediction, and PAM for identifying subgroups of A. baumannii patients with potentially distinct risk profiles.

Despite these advances, unsupervised clustering has not been applied to ICU patients with A. baumannii, nor has the integration of phenotype membership into supervised ML models been systematically studied. Addressing this gap, the present work applies a two-stage design to identify phenotypes and evaluate their added value for mortality prediction in critically ill patients with A. *baumannii* infection.

3. Methods

3.1 Study Design and Population

We worked on a previously conducted retrospective observational study of 231 intensive care unit (ICU) patients with confirmed A. *baumannii* infection, admitted to three different tertiary hospitals from the north, middle, and south of the West Bank, Palestine. Clinical and microbiological data were collected from hospital records across the three tertiary hospitals. Variables are summarized in Table A1 in Appendix. The primary endpoint was all-cause in-hospital mortality at discharge.

All methods were carried out in accordance with relevant guidelines and regulations. The use of anonymized human data in this study was conducted in accordance with both local and international ethical principles, including the Declaration of Helsinki. The protocol involving the data source was approved by the Institutional Review Board (IRB) of An-Najah National University (approval number Mas. Feb. 2023/7, approved on February 5, 2023). The requirement for informed consent was waived by the IRB.

3.2 Data Preparation and Preprocessing

The initial dataset contained 231 patient records and 49 variables. The structured preprocessing pipeline, presented in Figure 1, was applied to ensure data quality, reduce redundancy, and retain clinically meaningful features for analysis.

From the initial cohort of 231 eligible ICU patients, 19 (8.2%) had at least one missing value in predictors required for clustering or prediction, leaving 212 complete cases for all modeling analyses. Given the modest fraction of missing data and the limited sample size, we used a complete-case approach rather than multiple imputation, which allows to build robust and interpretable models while keeping additional assumptions about the missing-data mechanism to a minimum. Further, ten variables judged a priori to be unrelated to the study objectives were excluded, and records with incomplete data were removed, reducing the dataset from 231 to 212 patients. The full list of excluded variables and the rationale for their exclusion is provided in Appendix Table A2. Additionally, categorical features were label-encoded and continuous predictors were converted into pre-specified ordinal categories, chosen using standard ICU practice and the empirical distribution in our cohort to form meaningful categories with adequate counts. The full set of cut-points for all discretized variables is reported in Appendix Table A3. To refine the dataset, we applied variance filtering to remove features with minimal variability, as near-zero variance predictors add complexity without improving discrimination [25]. Binary and categorical variables were label-encoded, and features with variance less than 0.01 were removed as near-constant. This conservative cut-off is consistent with recent medical ML work using 0.01

to filter near-zero-variance predictors [26,27], and led to the exclusion of variables such as urinary catheter, cancer, and effective antibiotic therapy. Accordingly, the dataset was reduced from 49 to 35 variables. Next, correlation filtering was conducted to minimize redundancy. Associations among variables were quantified using Cramér's V, which is a normalized measure of association [28]. Variables with correlation coefficients ≥ 0.60 were considered redundant. For each correlated pair, the clinically more relevant feature was retained. Because several predictors were originally continuous, we also examined Spearman correlations between the original continuous variables as a sensitivity check and it showed a very similar pattern. This confirmed that strongly correlated pairs had already been removed, and among the remaining predictors the highest correlation was observed between hospital length of stay and duration of antibiotic therapy (ρ ≈ 0.65), which we judged clinically distinct and therefore retained both. After this step, 28 variables remained. Finally, the variable set was reviewed in consultation with specialists in the domain to ensure clinical interpretability and relevance. Six additional variables judged redundant or less informative were excluded: categorize reason, hypertension, neurological disease, infection with multidrug-resistant strains, white blood cell count, and platelet count.

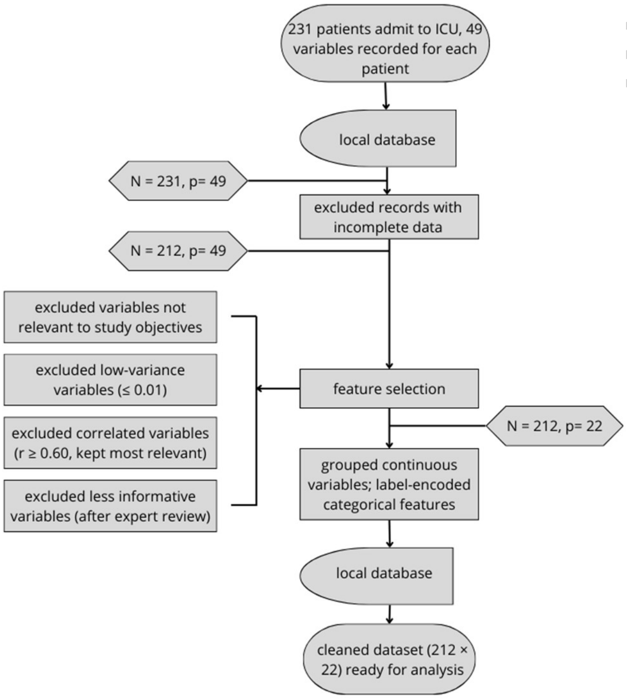
231 patients admit to ICU, 49 variables recorded for each patient
local database
N = 231, p= 49
excluded records with incomplete data
N = 212, p= 49
excluded variables not relevant to study objectives
excluded low-variance variables (≤ 0.01)
excluded correlated variables (r ≥ 0.60, kept most relevant)
excluded less informative variables (after expert review)
feature selection
N = 212, p= 22
grouped continuous variables; label-encoded categorical features
local database
cleaned dataset (212 × 22) ready for analysis

Figure 1: Preprocessing and feature selection workflow for the ICU dataset.

The final dataset comprised 212 patients and 22 clinically relevant variables, detailed in Table A3 in the Appendix, which were subsequently used for clustering and predictive modeling.

3.3 Clustering Analysis

Because ICU patients with A. *baumannii* infection are highly heterogeneous, we applied unsupervised clustering to identify subgroups of patients with similar clinical characteristics. This step aimed to reveal latent phenotypes that may underlie differences in prognosis and provide additional information to enhance supervised mortality prediction. Before clustering, dimensionality reduction was performed using Principal Component Analysis (PCA). PCA transforms correlated clinical variables into a smaller number of uncorrelated components while retaining most of the variability in the data. This reduces noise, mitigates the effect of collinearity, and improves the stability of clustering in high-dimensional datasets [29,30].

Clustering was then performed using the Partitioning Around Medoids (PAM) algorithm. Unlike centroid-based k-means, PAM selects actual patient records as cluster representatives “medoids” and minimizes the sum of dissimilarities to these medoids, which improves robustness to outliers and does not assume continuous variables with approximately spherical clusters [23]. We preferred PAM over agglomerative hierarchical clustering, which is sensitive to irreversible early merges and requires a subjective choice of dendrogram cut, and over density-based methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which require tuning density parameters and can leave some patients unassigned as noise, whereas our aim was to assign every ICU patient to a phenotype. Patient-to-medoid dissimilarities were computed using the Minkowski distance, a flexible metric that generalizes several distance measures and is applicable to our coded “binary/ordinal” clinical predictors [23].

The optimal number of clusters was evaluated using two complementary methods. The Elbow Method inspects the reduction in within-cluster variation and identifies the point where adding further clusters yields diminishing returns [31]. The Calinski–Harabasz Index compares the separation between clusters with cohesion within clusters, with higher values reflecting better-defined groupings [32].

To evaluate reproducibility, cluster stability was assessed using bootstrap resampling combined with the Jaccard similarity index, which quantifies how consistently patients are assigned to the same cluster across repeated resamples [24]. Finally, the cluster assignments derived from PAM were retained as an additional feature and incorporated into the supervised prediction models, as clarified in Section 3.5, to test whether phenotypic information improved mortality prediction.

3.4 Cluster Profiling

Following clustering, we profiled the subgroups to characterize their clinical and demographic features. For each cluster, mean and median values were computed to summarize age, comorbidities, infection characteristics, interventions, and outcomes.

To formally test for differences between clusters, non-parametric statistical methods were applied. Mann–Whitney U tests were used for ordinal variables, while chi-square tests were used for binary variables. For cluster profiling, nominal p-values were complemented by Benjamini–Hochberg False Discovery Rate (FDR–adjusted) q-values, where variables with $q < 0.05$ are highlighted as the most robust differences, and others are interpreted as exploratory. This approach follows prior phenotyping work in sepsis and ARDS, where cluster profiles were statistically validated to confirm their clinical distinctiveness [7,9]. Additionally, this approach provided interpretable phenotypes and allowed us to assess whether mortality differed significantly between clusters.

3.5 Supervised Prediction Models

To assess whether cluster membership improved prediction, we developed supervised machine-learning models. Cluster assignments obtained in Section 3.3 were incorporated as an additional feature alongside demographic, clinical, and infection-related variables. Two algorithms were selected based on their strong performance in structured clinical prediction tasks, RF and XGBoost [36].

RF is an ensemble method that constructs multiple decision trees using bootstrapped samples and aggregates their outputs. This approach reduces overfitting, captures nonlinear relationships, and produces measures of variable importance, making it valuable in clinical prognostic modeling [33]. XGBoost is a gradient boosting algorithm that builds trees sequentially, with each new tree correcting the errors of the previous ones. It incorporates regularization to prevent overfitting and has consistently shown high predictive accuracy in biomedical applications [34].

To examine the role of cluster membership across different clinical data conditions, we adopted a dual-model design. Each algorithm was trained under two complementary feature regimes: (1) a full-feature model including all 22 predictors, and (2) a reduced-feature model restricted to variables that differed significantly between clusters in Section 3.4. This design tested whether the cluster label contributes as an additional signal in data-rich settings or as a summary interaction feature in data-limited settings. Accordingly, for each algorithm, this yielded four configurations: a full-feature baseline model (all predictors, no cluster label), a full-feature cluster-enhanced model (all predictors plus the binary cluster label), a reduced-feature baseline model (reduced predictor set without the cluster label), and a reduced-feature cluster-enhanced model (the same reduced set plus the cluster label).

Model training and validation were conducted using 10-fold cross-validation, in order to provide stable performance estimates while minimizing overfitting [35]. For Random Forest and XGBoost models, the predictors were the original clinical variables, with or without the discrete phenotype label obtained from the PCA–PAM step. In each fold, the model was fitted on the training data and evaluated on the held-out fold. The mortality outcome in test folds was never used for phenotyping or model fitting, and the PCA components themselves were not entered into the supervised models; the cluster labels were treated as fixed baseline covariates.

Figure 2 summarizes the full methodological pipeline, including dimensionality reduction, clustering, and construction of baseline and enhanced supervised prediction models across the full feature set. The same process was also applied to reduced-feature models, restricted to variables that differed significantly between clusters.

This pipeline corresponds to the proposed phenotype-guided modeling framework, defined as a model-agnostic three-step pipeline that (i) derives data-driven phenotypes, (ii) integrates them as candidate features in supervised models, and (iii) formally evaluates their incremental predictive value. Additionally, it can in principle be applied to other ICU cohorts using different learning algorithms.

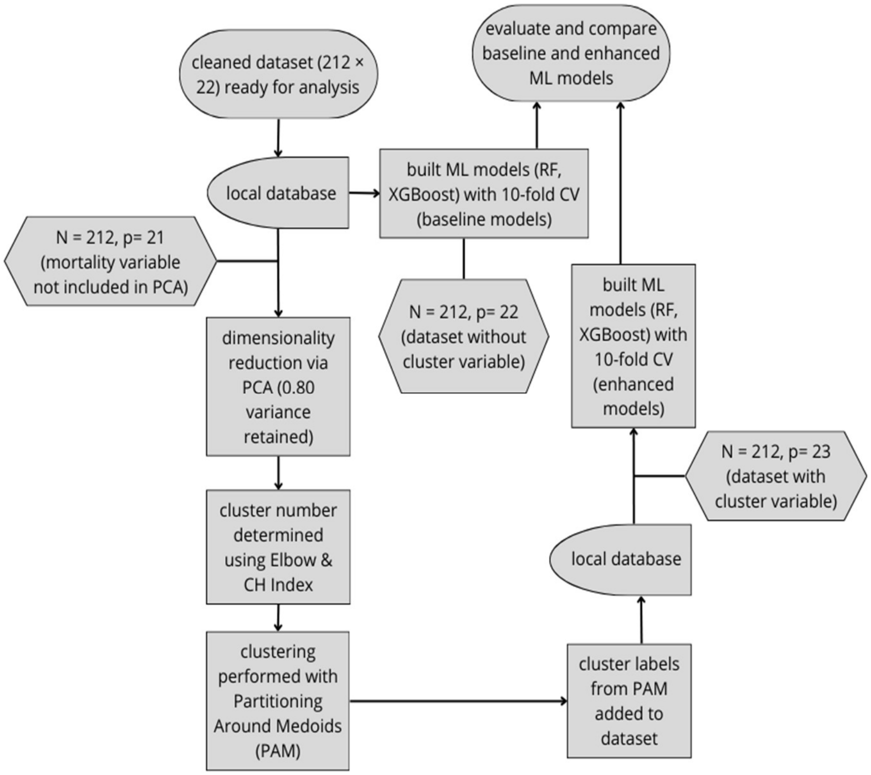
cleaned dataset (212 × 22) ready for analysis
local database
N = 212, p= 21 (mortality variable not included in PCA)
dimensionality reduction via PCA (0.80 variance retained)
cluster number determined using Elbow & CH Index
clustering performed with Partitioning Around Medoids (PAM)
built ML models (RF, XGBoost) with 10-fold CV (baseline models)
N = 212, p= 22 (dataset without cluster variable)
evaluate and compare baseline and enhanced ML models
built ML models (RF, XGBoost) with 10-fold CV (enhanced models)
N = 212, p= 23 (dataset with cluster variable)
local database
cluster labels from PAM added to dataset

Figure 2: Workflow of clustering and supervised prediction modeling for the full feature set.

3.6 Model Evaluation and Interpretability

To assess the ability of the supervised models to predict mortality, we evaluated their performance using a set of complementary metrics which capture different aspects of classification performance and to ensure a fair evaluation of predictive ability in clinical practice.

The area under the receiver operating characteristic curve (AUROC) was used as the primary measure of discrimination. AUROC quantifies how well the model distinguishes between survivors and non-survivors across all thresholds, with 0.5 indicating no discrimination and 1.0 representing perfect separation:

$$AUROC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx$$

Where *TPR* is sensitivity and *FPR* is the false positive rate [37].

To complement AUROC, several threshold-dependent metrics were also reported:

- Accuracy: the proportion of correct classifications among all cases:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

While intuitive, accuracy alone can be misleading in imbalanced data [38].

- Sensitivity (Recall): the proportion of deaths correctly identified:

$$Sensitivity = TP/(TP + FN)$$

A key measure in clinical settings where missing high-risk patients has serious consequences [40].

- Specificity: the proportion of survivors correctly identified:

$$Specificity = TN/(TN + FP)$$

This helps ensure the model avoids excessive false alarms [40].

- Balanced Accuracy: the average of sensitivity and specificity, designed to provide a fairer estimate of performance under imbalanced outcomes:

$$BalancedAccuracy = (Sensitivity + Specificity)/2$$

Recommended for skewed class distributions [41].

- F1-Score: the harmonic mean of precision and sensitivity, balancing the ability to detect deaths with the risk of false positives:

$$F1 = 2 \times (Precision \times Sensitivity)/(Precision + Sensitivity)$$

where

$$AUROC = \int \square^{1} TPR(FPR^{-1}(x))dx$$

which is particularly informative when positive cases (mortality) are less frequent [38].

Because mortality was less common than survival in our dataset, using this combination of metrics provided a more balanced and clinically meaningful assessment of predictive performance. For Random Forest and XGBoost models, predicted probabilities of in-hospital mortality were converted to class labels using a fixed threshold of 0.5. Accordingly, sensitivity, specificity, balanced accuracy, and F1-score were computed from the resulting confusion matrices at this threshold.

Finally, to enhance interpretability, we used Shapley Additive Explanations (SHAP). SHAP values assign a contribution of each feature, including conventional clinical predictors and cluster membership, to individual predictions. This allowed us to both rank variables by their global influence and explain risk estimation at the patient level [39].

4. Results

4.1 Study Population and Baseline Characteristics

The final analysis was conducted on 212 ICU patients with complete data on all predictors used for clustering and mortality prediction. The data included both male and female patients with a wide age distribution. Comorbid conditions were common, with substantial proportions presenting with diabetes mellitus, heart failure, chronic kidney disease, chronic liver disease, and chronic respiratory disease. Prior exposure to antibiotics and previous ICU admission within the last 90 days were also frequently observed. The majority of infections were hospital-acquired, and a notable fraction of patients were colonized with A. *baumannii* at nasal or rectal sites. Interventions such as mechanical ventilation and central venous catheter insertion were prevalent, reflecting the severity of illness in this cohort. The Sequential Organ Failure Assessment (SOFA) score varied widely, with patients distributed across low, moderate, and high severity categories. Length of stay in both the ICU and hospital was heterogeneous, with some patients requiring prolonged admissions.

The primary outcome variable was mortality at hospital discharge, which served as the dependent endpoint for subsequent analyses. Baseline demographic and clinical characteristics of the study population are summarized in Table A4 in the Appendix.

4.2 Clustering Outcomes

The study applied unsupervised clustering to identify subgroups of patients with similar clinical characteristics, and to uncover latent phenotypes that might explain differential outcomes and provide additional information for enhancing supervised mortality prediction. Because the dataset included 22 clinically relevant and intercorrelated variables, , we first applied principal component analysis (PCA) for dimensionality reduction. Following standard recommendations in clinical data analysis, we retained the first eight principal components, which together explained 82% of the total variance [29,30]. Inspection of PCA loadings (Appendix Table A5) showed that first principal component (PC1) was mainly driven by longer hospital stay, longer duration of antibiotic therapy, older age, and higher SOFA score, while the second principal component (PC2) captured a related gradient combining age, SOFA, length of stay, antibiotic duration, and comorbidities; together, these components can be interpreted as latent axes of acute severity, treatment intensity, and baseline vulnerability.

Clustering was then performed using the PAM algorithm in this eight-dimensional PCA space, following the procedure described in Section 3.3, and the resulting cluster assignments were used for subsequent profiling and prediction.

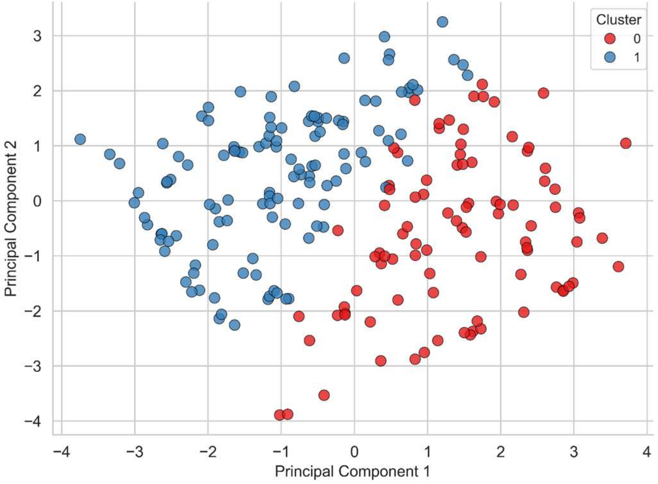

Figure 3. Distribution of ICU patients by cluster assignment on the first two principal components.

The optimal number of clusters was determined using complementary criteria applied to candidate solutions with k=2–10 (Appendix Fig. A1). The elbow plot of within-cluster sum of squares showed a marked decrease up to k=4 and only small improvements thereafter, while the Calinski–Harabasz index was maximal at k=2 and declined for larger k. Considering both criteria and this agreement between the two metrics, the moderate sample size, and the clinical interpretability of a simpler structure, we selected a two-cluster solution [29,30]. As a result, two clusters were identified, Cluster 0 (n = 85, 40%) and Cluster 1 (n = 127, 60%). The separation of the two clusters, projected onto the first two principal components, is illustrated in Figure 3. In this representation, Cluster 0 occupies the region with higher values on these severity and vulnerability components, consistent with its high-risk profile, whereas Cluster 1 is concentrated at lower values, matching the lower-risk phenotype described in Section 4.3. These clusters were then carried forward for clinical profiling and statistical comparison as presented in the following Section 4.3.

4.3 Cluster Profiling

The clinical and demographic features of the two clusters were characterized to determine whether they represented clinically meaningful phenotypes. Cluster profiling revealed marked differences in demographic, comorbidity, and outcome patterns. Cluster 0 represented a high-risk phenotype, characterized by older age, higher prevalence of heart failure, diabetes mellitus, chronic kidney and liver disease, higher SOFA scores, longer hospital stays, prolonged antibiotic therapy, and a mortality rate of 51%. In contrast, Cluster 1 represented a lower-risk phenotype, comprising younger patients with fewer comorbidities, lower SOFA scores, shorter hospital stays, shorter antibiotic courses, and a mortality rate of only 10%.

Table 1 summarizes the cluster comparison for the coded variables used in modeling, including p-values and FDR-adjusted q-values.. Out of the 21 clinical and demographic variables assessed, only nine showed nominal differences between clusters ($p < 0.05$), of which eight remained statistically significant after FDR correction ($q < 0.05$), where chronic liver disease variable retained a nominal association but did not pass the 5% FDR threshold and is therefore interpreted as exploratory. For completeness, the corresponding comparisons on the original continuous scale (age, SOFA score, length of stay, and duration of antibiotic therapy) are reported in Appendix Table A6, which confirms the same pattern of higher values in Cluster 0 (older age, higher SOFA, longer stays, longer antibiotic courses, and higher mortality) compared with Cluster 1. These results show that PAM clustering successfully identified two clinically meaningful phenotypes with distinct prognostic implications.

Table 1. Cluster comparison of key baseline variables (coded ordinal/binary predictors) with p-values and FDR-adjusted q-values.

Variable	**Type**	**P-value**	**FDR (q-value)**	**Cluster 0**		**Cluster 1**		**Test**
				Mean	**Median**	**Mean**	**Median**	
Age	Ordinal	< 0.05 (Significant	< 0.05 (Significant	0.96	1	2.57	3	Mann-Whitney U
Heart failure	Binary	< 0.05 (Significant	< 0.05 (Significant	0.24	0	0.65	1	Chi-square
Diabetes mellitus	Binary	< 0.05 (Significant	< 0.05 (Significant	0.48	0	0.7	1	Chi-square
Chronic kidney disease	Binary	< 0.05 (Significant	< 0.05 (Significant	0.12	0	0.28	0	Chi-square
Chronic liver disease	Binary	< 0.05 (Significant	> 0.05 (Not Significant	0.07	0	0.18	0	Chi-square
SOFA score	Ordinal	< 0.05 (Significant	< 0.05 (Significant	1.4	1.5	2.8	3	Mann-Whitney U
Length of hospital stay	Ordinal	< 0.05 (Significant	< 0.05 (Significant	2.9	3	1.6	1.25	Mann-Whitney U

Duration of antibiotic therapy	Ordinal	< 0.05 (Significant	< 0.05 (Significant	1.9	2	0.95	2	Mann-Whitney U
Mortality at discharge	Binary	< 0.05 (Significant	< 0.05 (Significant	0.51	1	0.1	0	Chi-square

4.4 Predictive Modeling

Supervised prediction models, RF and XGBoost, were trained to evaluate whether adding cluster membership improved mortality prediction under both full-feature and reduced-feature settings. For each configuration, point estimates of the evaluation metrics were obtained from 10-fold cross-validation (CV), and 95% confidence intervals (CIs) were derived by non-parametric bootstrapping (1,000 resamples) of the out-of-fold predictions.

Table 2. Random Forest performance (10-fold CV) with 95% bootstrap CIs for full and reduced models, with and without cluster membership.

Metric	**Full Model**		**Reduced model**	
	RF Baseline	**RF Cluster-Enhanced**	**RF Baseline**	**RF Cluster-Enhanced**
AUROC	0.95 (0.92 – 0.99)	0.95 (0.92 – 0.99)	0.92 (0.89 – 0.98)	0.93 (0.88 – 0.97)
Accuracy	0.91 (0.89 – 0.96)	0.93 (0.91 – 0.97)	0.91 (0.89 – 0.95)	0.91 (0.88 – 0.95)
Balanced Accuracy	0.87 (0.84 – 0.94)	0.90 (0.86 – 0.95)	0.88 (0.83 – 0.93)	0.88 (0.83 – 0.93)
Sensitivity	0.79 (0.71 – 0.91)	0.82 (0.76 – 0.93)	0.79 (0.71 – 0.90)	0.79 (0.70 – 0.90)
Specificity	0.96 (0.95 – 0.99)	0.97 (0.95 – 0.99)	0.95 (0.92 – 0.98)	0.95 (0.92 – 0.99)
F1-score	0.89 (0.86 – 0.95)	0.91 (0.88 – 0.96)	0.88 (0.84 – 0.93)	0.88 (0.84 – 0.94)
F1(Minority)	0.83 (0.78 – 0.92)	0.87 (0.82 – 0.94)	0.82 (0.76 – 0.90)	0.82 (0.75 – 0.90)

4.4.1 Random Forest Models

The RF full-feature baseline model achieved strong discriminative performance. As seen in Table 2, adding cluster membership improved several metrics: balanced accuracy increased to 0.90, F1(Minority) to 0.87, and sensitivity rose from 0.79 to 0.82, while AUROC remained stable at 0.95. These improvements indicate that cluster membership contributed additional prognostic signal beyond conventional clinical predictors, enhancing the identification of high-risk patients. In contrast, when using the reduced feature set, RF performance remained unchanged with or without cluster membership. This suggests that, for RF, the reduced set of statistically significant variables already captured the most discriminative information, leaving little added value from cluster membership. Generally, RF results highlight that phenotype information is most valuable in data-rich settings, where it acts as an additional predictor to improve sensitivity and balanced classification.

4.4.2 XGBoost Models

The opposite pattern was observed with XGBoost. For the full-feature baseline model, it also demonstrated high performance, with AUROC 0.95 and balanced accuracy 0.90. Inclusion of cluster membership produced no measurable improvement, as all metrics remained essentially unchanged as presented in Table 3. This indicates that the full feature set was sufficient for XGBoost to capture most of the prognostic information, and the cluster label did not add incremental benefit. This reflects that XGBoost was already capturing the relevant interactions among predictors. By contrast, in the reduced-feature setting, cluster membership provided clear gains. Balanced accuracy increased from 0.86 to 0.89, sensitivity from 0.79 to 0.84, and the F1-score for the minority (mortality) class from 0.81 to 0.84. These results demonstrate that when fewer variables were available, the cluster label acted as a compact summary of underlying interactions, thereby restoring predictive capacity.

Table 3. XGBoost performance (10-fold CV) with 95% bootstrap CIs for full and reduced models, with and without cluster membership.

Metric	Full Model		Reduced model	
	XGBoost Baseline	XGBoost Cluster-Enhanced	XGBoost Baseline	XGBoost Cluster-Enhanced
AUROC	0.95 (0.91 – 0.98)	0.95 (0.91 – 0.98)	0.95 (0.90 – 0.97)	0.94 (0.90 – 0.97)
Accuracy	0.92 (0.88 – 0.95)	0.93 (0.90 – 0.91)	0.90 (0.86 – 0.93)	0.92 (0.89 – 0.96)
Balanced Accuracy	0.89 (0.85 – 0.94)	0.91 (0.86 – 0.95)	0.87 (0.81 – 0.92)	0.90 (0.86 – 0.95)
Sensitivity	0.84 (0.75 – 0.93)	0.86 (0.78 – 0.94)	0.79 (0.69 – 0.89)	0.86 (0.77 – 0.94)
Specificity	0.95 (0.91 – 0.98)	0.95 (0.92 – 0.99)	0.94 (0.90 – 0.97)	0.95 (0.91 – 0.98)
F1-score	0.89 (0.85 – 0.94)	0.91 (0.87 – 0.95)	0.87 (0.82 – 0.92)	0.90 (0.86 – 0.95)
F1(Minority)	0.84 (0.78 – 0.92)	0.87 (0.80 – 0.93)	0.81 (0.74 – 0.88)	0.86 (0.79 – 0.92

4.5 Feature Importance Analysis

To further interpret model predictions, we examined feature importance rankings from the RF and XGBoost classifiers as presented in Figures 4 and 5. In all configurations, the SOFA score emerged as the strongest predictor of mortality, clearly outweighing all other variables. Additional influential features included mechanical ventilation, chronic kidney disease, length of hospital stay, and age, while infection-related variables such as hospital-acquired infection and *A. baumannii* colonization consistently had negligible influence on mortality prediction.

In the RF full model, cluster membership appeared as one of the leading predictors, positioned just below SOFA score and other severity indicators as seen in Figure 4, which is consistent with the observed improvement in model performance when cluster was added. In contrast, according to Figure 5, XGBoost in the full feature setting assigned only moderate weight to cluster membership, placing greater emphasis on conventional severity and comorbidity variables, which aligns with the minimal performance change reported for this model.

When analyses were restricted to the nine variables that differed significantly between clusters, RF continued to assign cluster a measurable role, but its relative importance declined compared to the full model, reflecting the lack of additional predictive gain. XGBoost, however, increased the relative weight of cluster membership in the reduced model, ranking it among mid-level predictors and producing measurable improvements in balanced accuracy and sensitivity.

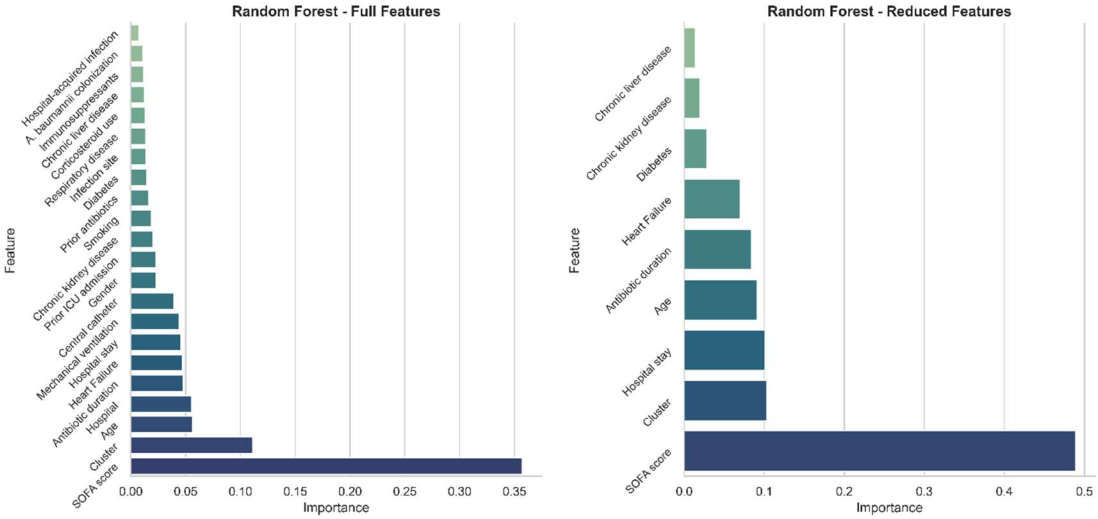

Figure 4. Feature importance ranking for mortality prediction using RF with full feature (left) and reduced feature (right) sets.

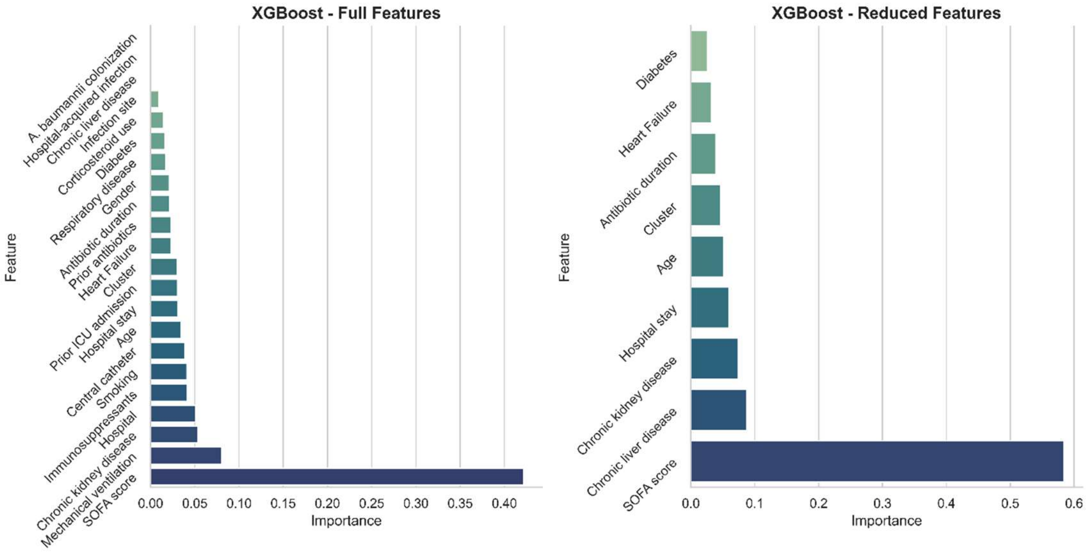

Figure 5. Feature importance ranking for mortality prediction using XGBoost with full feature (left) and reduced feature (right) sets.

These complementary patterns confirm that while severity measures dominate mortality prediction across algorithms, the phenotypes derived from clustering can provide meaningful additional signal, with their influence depending on the modeling strategy and the feature space considered.

5. Discussion and Conclusions

In this study, we applied a hybrid "unsupervised-supervised" ML pipeline to ICU patients with A. *baumannii* infection and demonstrated that clustering revealed two reproducible phenotypes with distinct clinical characteristics and mortality risks. This extends prior work in sepsis and ARDS [7-9,14-16], showing that machine-learned phenotypes can uncover heterogeneity often masked in conventional models of A. *baumannii* . Because all predictors are obtainable during the ICU stay, these mortality predictions are intended to support early identification of high-risk patients and to inform decisions about monitoring, antimicrobial optimization, organ support, and goals of care.

Supervised analyses confirmed that phenotypes add predictive value, but our results also show that this enhancement is not universal. In RF, cluster membership contributed additional prognostic signal in the full model, ranking just below SOFA score in importance and improving discrimination beyond conventional variables. Yet, when only reduced variables were included, the added value of cluster membership diminished, since the most discriminative features had already been retained. By contrast, in XGBoost, cluster membership provided little benefit in the full model but became more influential in the reduced setting, acting as a summary marker that compensated for the smaller input space. This demonstrates an important methodological contribution: the utility of phenotype labels depends not only on the data but also on the interaction between algorithm choice and feature regime. To our knowledge, this is the first study to explicitly test this “algorithm × feature regime” interaction in a clinical ML context, highlighting a contribution that is relevant beyond medicine.

Feature importance analyses provided further context. Across models, SOFA score consistently emerged as the strongest predictor of mortality, followed by mechanical ventilation, chronic kidney disease, and age, confirming established knowledge that severity of organ dysfunction and invasive support are central determinants of outcome in A. *baumannii* infection [3,4,12]. Cluster membership, while not the dominant predictor, provided complementary signal under specific circumstances, underscoring its role as an additional, not universal, contributor to prognostic modeling.

From a clinical perspective, phenotype-guided models are best seen as tools for early risk stratification rather than as deterministic prognostic scores. Using routine ICU data, high-risk phenotypes could be flagged for closer monitoring, optimization of antimicrobial therapy and source control, and escalation of organ support, while lower-risk phenotypes may support more cautious de-escalation under clinical judgment. Because phenotypes improved performance only in specific algorithm–feature regimes, they are best viewed as risk modifiers that complement existing severity scores and help design trials or quality-improvement efforts in clearly defined high-risk subgroups.

From an ML perspective, this dual design illustrates two distinct roles of unsupervised phenotypes: as an extra layer of information in data-rich models (RF, full set) and as a compact summary of interactions in data-limited settings (XGB, reduced set). This clarifies why clustering sometimes improves prediction and sometimes does not, resolving a common ambiguity in the literature.

This study has some considerations when interpreting the findings. It was conducted retrospectively across three hospitals, which provides valuable real-world data but may not capture the full spectrum of patient populations. Additionally, missing predictors were handled using complete-case analysis, which assumes that missingness does not depend on unobserved outcomes or covariates. In a retrospective ICU chart review, missing values are likely related to documentation and ordering patterns and thus approximately Missing at Random (MAR), but some deviation from this assumption cannot be excluded. As a result, our estimates may be subject to residual bias and the effective sample size for model development is slightly reduced. Moreover, model performance was assessed using internal cross-validation, which is a robust approach for internal validation, though future prospective and multi-center studies will be important to confirm reproducibility in broader contexts. The clustering results may also vary depending on the input variables and distance metric, which is inherent to most unsupervised methods and underscores the need for methodological comparisons in future research. The dataset was also imbalanced, with mortality more frequent than survival. While balanced accuracy and F1 scores were used to mitigate bias, larger and more diverse datasets or advanced resampling techniques would help confirm the stability of the observed effects. In this methodological study, we used a fixed 0.5 threshold for all models to allow fair comparison of algorithm–feature regimes; in future work, threshold optimisation based on Youden’s J or explicit clinical trade-offs could be explored for deployment in practice. Moreover, clustering was performed once on the full cohort using baseline clinical variables, and the resulting phenotypes were treated as fixed labels in the prediction models. This two-stage approach is standard in ICU phenotyping studies, where phenotypes are derived once from baseline data and then used to study prognosis or treatment response [7,9,14–16]. From a strict predictive-pipeline perspective, a fully automated deployment system would refit PCA and clustering within training samples and assign phenotypes to new patients; this is a natural refinement for future external validation work. Finally, as with most retrospective studies, some unmeasured clinical factors were not available, highlighting the value of integrating richer data sources in future work.

Acknowledgments: The authors would like to thank An-Najah National University (www.najah.edu) for the technical support provided to publish the present manuscript.

Author Contributions: A.E. and R.K. contributed conceptualization; A.E. and S.A. contributed methodology; A.E, R.K. and H.R. contributed validation; A.E. and S.A. contributed formal analysis and writing; R.K. and H.R. contributed writing, review and editing. The published version of the manuscript has been read and approved by all authors.

Data Availability Statement: The corresponding author has access to the data used to support the conclusions of this investigation and it will be available on reasonable request.

Funding: Not applicable

Informed Consent Statement: Informed consent was waived by the Institutional Review Board (IRB) of An-Najah National University (Mas. Feb. 2023/7, February 5, 2023).

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

References

1 Moubareck C, Hammoudi Halat D. Insights into *Acinetobacter baumannii*: a review of microbiological, virulence, and resistance traits in a threatening nosocomial pathogen. *Antibiotics (Basel).* 2020;9(3):119. doi:10.3390/antibiotics9030119.

2 World Health Organization. Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis. Geneva: WHO; 2017.

3 Falagas ME, Bliziotis IA, Siempos II. Attributable mortality of *Acinetobacter baumannii* infections in critically ill patients: a systematic review of matched cohort and case–control studies. *Crit Care.* 2006;10(2):R48. doi:10.1186/cc4869.

4 Itani R, Araj GF, Kanj SS, Dbaibo G, Kanj NA, Jaber F, et al. *Acinetobacter baumannii*: assessing susceptibility patterns, management practices, and mortality predictors in a tertiary teaching hospital in Lebanon. *Antimicrob Resist Infect Control.* 2023;12:136. doi:10.1186/s13756-023-01343-8.

5 Xu J, Chen X, Zheng X, Tong K, Liu J, Liu Y, et al. *Acinetobacter baumannii* complex–caused bloodstream infection in ICU during a 12-year period: predicting fulminant sepsis by interpretable machine learning. *Front Microbiol.* 2022;13:1037735. doi:10.3389/fmicb.2022.1037735.

6 Zhang G, Shao F, Yuan W, et al. Predicting sepsis in-hospital mortality with machine learning: a multi-center study using clinical and inflammatory biomarkers. *Eur J Med Res.* 2024;29:156. doi:10.1186/s40001-024-01756-0.

7 Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA.* 2019;321(20):2003–2017. doi:10.1001/jama.2019.5791.

8 Aldewereld ZT, Zhang LA, Urbano A, Liu W, Yang J, Choi J, et al. Identification of clinical phenotypes in septic patients presenting with hypotension or elevated lactate. *Front Med.* 2022;9:794423. doi:10.3389/fmed.2022.794423.

9 Calfee CS, Delucchi KL, Parsons PE, Thompson BT, Ware LB, Matthay MA. Subphenotypes in acute respiratory distress syndrome: latent class analysis of ARDS clinical trials. *Lancet Respir Med.* 2014;2(8):611–20. doi:10.1016/S2213-2600(14)70097-9.

10 Tang L, Xu S, Li H, Wang Q, Chen Y, Guo Z, et al. Machine learning model to predict sepsis in ICU patients with intracerebral hemorrhage. *Sci Rep.* 2025;15:99431. doi:10.1038/s41598-025-99431-9.

11 Özdede M, Zarakolu P, Metan G, et al. Predictive modeling of mortality in carbapenem-resistant *Acinetobacter baumannii* bloodstream infections using machine learning. *J Investig Med.* 2024;72(7):684–96. doi:10.1177/10815589241258964.

12 Song H, Zhang H, Zhang D, et al. Establishment and validation of a risk prediction model for mortality in patients with *Acinetobacter baumannii* infection: a retrospective study. *Infect Drug Resist.* 2023;16:7855–66. doi:10.2147/IDR.S423969.

13 Neuman I, Shvartser L, Teppler S, et al. A machine-learning model for prediction of *Acinetobacter baumannii* hospital-acquired infection. *PLoS One.* 2024;19(12):e0311576. doi:10.1371/journal.pone.0311576.

14 Bai Y, Xia J, Huang X, Chen S, Zhan Q. Using machine learning for the early prediction of sepsis-associated ARDS in the ICU and identification of clinical phenotypes with differential responses to treatment. *Front Physiol.* 2022;13:1050849. doi:10.3389/fphys.2022.1050849.

15 Sinha P, Churpek MM, Calfee CS. Machine learning classifier models can identify ARDS phenotypes using readily available clinical data. *Am J Respir Crit Care Med.* 2020;202(7):996–1004. doi:10.1164/rccm.202002-0347OC.

16 Maddali MV, Sinha P, Walkey AJ, et al. Validation and utility of ARDS subphenotypes identified by machine learning using clinical data. *Lancet Respir Med.* 2022;10(4):367–77. doi:10.1016/S2213-2600(21)00461-6.

17 Su C, Zhang Y, Flory JH, et al. Clinical subphenotypes in COVID-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health. *NPJ Digit Med.* 2021;4:110. doi:10.1038/s41746-021-00481-w.

18 Velez T, Song K, Gray K, et al. Identification and prediction of clinical phenotypes in hospitalized patients with COVID-19. *JMIR Form Res.* 2023;7(1):e46807. doi:10.2196/46807.

19 Yamga E, d'Alessandro AF, Jutant E-M, et al. Interpretable clinical phenotypes among patients hospitalized with COVID-19 using cluster analysis. *Front Digit Health.* 2023;5:1142822. doi:10.3389/fdgth.2023.1142822.

20 Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the ICU using machine learning and physiological data. *Biomed Eng Online.* 2016;15(Suppl 1):124. doi:10.1186/s12938-016-0230-8.

21 Nemati S, Holder A, Razmi F, et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med.* 2018;46(4):547–53. doi:10.1097/CCM.0000000000002936.

22 Johnson AEW, Ghassemi MM, Nemati S, et al. Machine learning and decision support in critical care. *Proc IEEE.* 2016;104(2):444–66. doi:10.1109/JPROC.2015.2501978.

23 Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley; 2009.

24 Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal.* 2007;52(1):258–71. doi:10.1016/j.csda.2006.11.025.

25 Kuhn M, Johnson K. *Applied Predictive Modeling.* New York: Springer; 2013. doi:10.1007/978-1-4614-6849-3.

26 Castillo T JM, Starmans MPA, Arif M, Niessen WJ, Klein S, Bangma CH, Schoots IG, Veenland JF. *A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: high grade vs. low grade.* Diagnostics. 2021;11(2):369. doi: 10.3390/diagnostics11020369

27 Vy ND, Nguyen HT, Tran TT, Ngo DT, Pham DT, Ly NTT, et al. *Development of machine learning models using MRI radiomics and clinical features to predict lymphovascular space invasion and deep myometrial invasion in endometrial cancer.* American Journal of Cancer Research. 2024;14(6):2621–2645. doi: 10.7150/ajcr.90138

28 Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med.* 2018;18(3):91–93. doi:10.1016/j.tjem.2018.08.001.

29 Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Phil Trans R Soc A.* 2016;374:20150202. doi:10.1098/rsta.2015.0202.

30 Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat.* 2010;2(4):433–59. doi:10.1002/wics.101.

31 Ketchen DJ, Shook CL. The application of cluster analysis in strategic management research: an analysis and critique. *Strateg Manag J.* 1996;17(6):441–58. doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

32 Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Theory Methods.* 1974;3(1):1–27. doi:10.1080/03610927408827101.

33 Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. doi:10.1023/A:1010933404324.

34 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proc 22nd ACM SIGKDD Int Conf Knowledge Discovery and Data Mining.* 2016:785–94. doi:10.1145/2939672.2939785.

35 Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc 14th Int Joint Conf Artificial Intelligence.* 1995;2:1137–45.

36 Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart.* 2018;104(14):1156–64. doi:10.1136/heartjnl-2017-311198.

37 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29–36. doi:10.1148/radiology.143.1.7063747.

38 Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21:6. doi:10.1186/s12864-019-6413-7.

39 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.* 2017;30:4765–74.

40 Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* 2006;27(8):861–74. doi:10.1016/j.patrec.2005.10.010.

41 Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. *Proc 20th Int Conf Pattern Recognit.* 2010:3121–4. doi:10.1109/ICPR.2010.764.