April – June, 2026
Original Article
DOI: http://dx.doi.org/10.30867/action.v11i2.3304

AcTion: Aceh Nutrition Journal
Pages: 811 – 825 | p-issn 2527-3310; e-issn 2548-5741

Explainable Stacked Ensemble Machine Learning for predicting monthly Gross Death Rate using aggregated clinical and operational hospital indicators in Indonesia

Explainable Stacked Ensemble Machine Learning untuk memprediksi Gross Death Rate bulanan berdasarkan indikator klinis dan operasional rumah sakit teragregasi di Indonesia

Sri Murdiati[1*], Safrizal Rahman[2], Yoga Yuniadi[3], Nirwana Lazuardi Sary[4]

[1] Universitas Syiah Kuala, Banda Aceh, Indonesia. E-mail: sri.murdiati@usk.ac.id
[2] Universitas Syiah Kuala, Banda Aceh, Indonesia. E-mail: safrizalrahman@usk.ac.id
[3] Universitas Indonesia, Jakarta, Indonesia. E-mail: yogay136@gmail.com
[4] Universitas Syiah Kuala, Banda Aceh, Indonesia. E-mail: nirwana@usk.ac.id

***Correspondence Author:**
Department of Cardiology and Vascular Medicine, Universitas Syiah Kuala, Kopelma Darussalam, Banda Aceh 23111 Aceh, Indonesia.
E-mail: sri.murdiati@usk.ac.id

Article History:
Received: June 09, 2026; Revised: June 17, 2026; Accepted: June 24, 2026; Published: June 26, 2026.

Publisher:
Politeknik Kesehatan Aceh
Kementerian Kesehatan RI

Abstract

Accurate hospital-wide mortality prediction is important for institutional clinical governance, quality improvement and resource planning. However, most machine learning studies have focused on patient-level data, intensive care populations, or disease-specific cohorts, with limited integration of hospital operational indicators. This retrospective single-center predictive modeling study used 36 monthly institutional observations from January 2022 to December 2024 at a regional referral hospital in Indonesia. Aggregated clinical severity indicators were combined with operational performance metrics, including length of stay, bed occupancy rate, and bed turnover rates. Random Forest, XGBoost, and feedforward neural network models were developed and compared with a linear regression baseline. The performance was internally evaluated using time-aware five-fold cross-validation with R^2, root mean squared error (RMSE), and mean absolute error (MAE). GDR and error metrics were expressed as deaths per 1,000 admissions. The stacked ensemble achieved the highest R^2 (0.841) and the lowest RMSE (4.49 deaths per 1,000 admissions), while the neural network achieved the lowest MAE (2.74 deaths per 1,000 admissions). In conclusion, operational indicators had modest direct effects but improved the model robustness through interaction effects. These findings support the use of explainable ensemble machine learning for institutional-level mortality prediction and hospital decision support.

Keywords: Gross Death Rate, hospital mortality, hospital operational indicators, machine learning, stacked ensemble, SHAP

Abstrak

Prediksi mortalitas tingkat rumah sakit yang akurat penting untuk tata kelola klinis institusional, peningkatan mutu, dan perencanaan sumber daya. Namun, sebagian besar studi machine learning berfokus pada data tingkat pasien, populasi perawatan intensif, atau kohort spesifik penyakit, dengan integrasi indikator operasional rumah sakit yang masih terbatas. Studi pemodelan prediktif retrospektif satu pusat ini menggunakan 36 observasi institusional bulanan dari Januari 2022 hingga Desember 2024 pada rumah sakit rujukan regional di Indonesia. Indikator keparahan klinis teragregasi dikombinasikan dengan metrik kinerja operasional, termasuk length of stay, bed occupancy rate, dan bed turnover rate. Model Random Forest, XGBoost, dan feed-forward neural network dikembangkan dan dibandingkan dengan baseline regresi linear. Kinerja dievaluasi secara internal menggunakan time-aware five-fold cross-validation dengan R^2, root mean squared error (RMSE), dan mean absolute error (MAE). GDR dan metrik kesalahan dinyatakan sebagai kematian per

1.000 admisi. Stacked ensemble mencapai R^2 tertinggi (0,841) dan RMSE terendah (4,49 kematian per 1.000 admisi), sedangkan neural network mencapai MAE terendah (2,74 kematian per 1.000 admisi). Kesimpulan, indikator operasional memiliki efek langsung moderat, tetapi meningkatkan robustness model melalui efek interaksi. Temuan ini mendukung penggunaan explainable ensemble machine learning untuk prediksi mortalitas tingkat institusi dan decision support rumah sakit.

Kata Kunci: *Gross Death Rate, indikator operasional rumah sakit, machine learning, mortalitas rumah sakit, stacked ensemble, SHAP*

Introduction

Machine learning (ML) is increasingly used to predict hospital mortality and support risk surveillance, quality improvement, and clinical governance in hospitals. However, most existing studies have focused on individual patient outcomes using electronic health records (EHRs), intensive care unit (ICU) cohorts, or disease-specific samples. Although these patient-level models are valuable for bedside risk stratification, they differ conceptually and operationally from institutional-level models designed to estimate the monthly hospital mortality burden for quality dashboards, resource allocation, and service planning (König et al., 2022). Recent studies suggest that ML methods, including ensemble learning, deep neural networks, and decision tree–based algorithms, may outperform conventional statistical approaches in selected clinical settings. For example, decision tree models have achieved area under the receiver operating characteristic curve values approaching 0.96 in ICU populations, whereas boosting-based ensemble classifiers have demonstrated better predictive performance than regression-based models in specific patient groups, including patients with heart failure (Jawadi et al., 2024; Trentino et al., 2022; Yun et al., 2021).

Despite these advances, concerns remain regarding model generalizability, interpretability, and external validation, which limit routine clinical implementation. Therefore, a broader, multifactorial approach has emerged, in which clinical determinants are combined with operational and organizational information. Evidence indicates that incorporating structured admission data may improve mortality risk estimation, whereas real-time EHR integration may enable hospital-wide predictive monitoring (Deschepper et al., 2020; Giwangkancana et al., 2024; C. Li et al., 2021; Nie et al., 2021; Shah et al., 2020; Thorsen-Meyer et al., 2020; Wu & Gao, 2023).

Nevertheless, this study has several limitations. First, many mortality prediction models rely heavily on administrative variables and insufficiently incorporate relevant biomarkers, physiological measurements, and hospital operational indicators into their design. This restricted variable coverage may reduce the predictive performance and limit the transportability of the model across populations and institutional settings (Huang et al., 2023; Naemi et al., 2021). Second, the interpretability of existing models varies significantly. The black-box characteristics of advanced ML algorithms may reduce transparency, undermine clinician trust, and hinder their integration into routine workflows (Deng et al., 2023; Maheswari et al., 2023). Third, much of the evidence is derived from ICU populations or single-center datasets, whereas hospital-wide operational performance measures are rarely included in the predictive frameworks. Model performance may also deteriorate over time without systematic monitoring, recalibration, and maintenance (C. Li et al., 2021). Fourth, external validity is frequently insufficiently examined because cross-institutional validation and adaptation for high-risk populations, including pediatric, oncology, and other specialized patient groups, remain limited (B. Lee et al., 2021; Qiao et al., 2022; Seki et al., 2021; Sinha et al., 2024; Warman et al., 2022). These limitations highlight the need for interpretable, operationally relevant, institution-wide approaches that extend beyond narrowly defined patient-level or ICU-based frameworks.

Existing hospital mortality models also tend to underrepresent enterprise-level performance measures, reducing their usefulness in hospital management and operational decision-making. In clinical practice, institutional mortality patterns may be

associated not only with patient severity but also with bed availability, patient flow, ward workload, and service capacities. Nevertheless, key operational measures such as the bed occupancy rate (BOR), length of stay (LOS), and bed turnover rate (BTR) are seldom integrated with clinical covariates in ML-based mortality prediction models (Fang et al., 2022; Kwun et al., 2025; S. W. Lee et al., 2022; L. Li et al., 2023; M. Li et al., 2024; Muralitharan et al., 2021; Thorsen-Meyer et al., 2020). Therefore, an explainable ML framework may be useful for examining whether aggregated clinical and operational indicators can predict the monthly institutional mortality burden. However, because the present study used only 36 monthly observations from a single hospital, it was designed as an exploratory, internally evaluated analysis rather than an externally validated clinical decision support system.

This study contributes to the literature by linking healthcare operations with institutional mortality prediction and addressing the predominantly ICU-centered orientation of previous research (M. Li et al., 2024). Methodologically, it applies a reproducible ML and explainable artificial intelligence workflow for a transparent, institution-level performance evaluation (Fang et al., 2022; Lv et al., 2021). From an operational perspective, integrating clinical and hospital performance indicators may support mortality surveillance, capacity planning, staffing decisions, and the development of quality management dashboards (Estiri et al., 2021; S. W. Lee et al., 2022). The proposed framework may also provide a basis for future external validation, prospective implementation, temporal monitoring, and fairness assessment across heterogeneous hospital settings (Halasz et al., 2021; Hsu et al., 2020).

This study aimed to develop and internally evaluate explainable ML models for predicting the monthly gross death rate using aggregated clinical severity indicators and hospital operational performance measures from regional referral hospitals. Explainable artificial intelligence was incorporated using Shapley additive explanations (SHAP) to improve model transparency and facilitate the interpretation of hospital-wide mortality monitoring. Specifically, this study compared ensemble ML models with individual ML algorithms and linear regression, examined whether integrating BOR, LOS, and BTR with clinical covariates improved prediction, assessed the relative contributions of clinical and operational indicators, and evaluated whether interaction terms among operational measures improved predictive performance and model stability.

Methods

Study Design and Analytical Framework

This retrospective, single-center, predictive modeling study used aggregated monthly institutional data. No interventions, randomization, or experimental allocations were performed. The unit of analysis was the hospital-month, and the primary outcome was the monthly Gross Death Rate (GDR), expressed as the number of in-hospital deaths per 1,000 admissions.

The analytical framework compared a conventional statistical baseline model with individual machine learning models and a stacked ensemble model. Linear regression was used as the baseline model for this study. The individual machine learning models were random forest, extreme gradient boosting (XGBoost), and multilayer perceptron (MLP) neural networks. The stacked ensemble incorporated random forest and XGBoost as base learners, and a multilayer perceptron as the meta-learner. These models were selected to compare a directly interpretable linear approach with algorithms capable of representing nonlinear associations and interactions among institutional, clinical, and operational indicators.

Data Source and Study Setting

Data were obtained from Rumah Sakit Umum Daerah Dr. Zainoel Abidin, a regional tertiary referral hospital in Banda Aceh, Indonesia. The hospital receives patients from multiple districts in Aceh Province. The dataset consisted of aggregated monthly hospital activity, mortality, clinical severity, and operational performance records collected over 36 months from January 2022 to December 2024.

During the observation period, the hospital recorded 42,685 admissions and 1,142 in-hospital deaths. The dataset includes 36 monthly observations and 18 candidate predictor variables. These predictors comprised hospital operational indicators and aggregated measures of clinical severity. The operational

indicators included the average length of stay (LOS), bed occupancy rate (BOR), bed turnover rate (BTR), and bed turnover interval (TOI). The aggregated clinical indicators included heart failure, cardiogenic shock, and deep vein thrombosis (DVT). The clinical variables represented monthly institutional indicators of disease burden, clinical severity, and case mix, rather than individual patient-level characteristics.

The principal dataset characteristics are listed in Table 1. GDR and Net Death Rate (NDR) were expressed as deaths per 1,000 admissions to maintain consistency among descriptive statistics, model performance measures, and graphical displays.

Table 1. Characteristics of the institutional dataset

Characteristic	Value
Hospital type	Regional referral hospital
Observation period	January 2022 – December 2024
Total observation units	36 monthly records
Total hospital admissions	42,685
Total in-hospital deaths	1,142
Average monthly admissions	1,186
Gross Death Rate (mean)	26.7 deaths per 1,000 admissions
Gross Death Rate (range)	19.0-35.2 deaths per 1,000 admissions
Net Death Rate (mean)	18.4 deaths per 1,000 admissions
Average Length of Stay (LOS)	5.7 days
Bed Occupancy Rate (BOR)	73.4 %
Bed Turnover Rate (BTR)	6.2 times/month
Average Bed Turnover Interval (TOI)	1.9 days
Number of predictor variables	18
Outcome variable	Monthly Gross Death Rate (GDR), deaths per 1,000 admissions

Outcome and Predictor Variables

The primary outcome was the monthly GDR, defined as the number of in-hospital deaths occurring during a calendar month divided by the total number of hospital admissions during that month and multiplied by 1,000. GDR was analyzed as a continuous institutional-level outcome representing temporal variation in the hospital-wide mortality burden. Accordingly, the model estimated mortality at the hospital-month level and did not estimate the probability of death in individual patients. Therefore, the associations between the predictors and GDR were interpreted only at the institutional level to avoid ecological inference at the patient level.

The predictors were grouped into operational, clinical, and mortality-derived contextual indicators. Operational indicators describe hospital capacity and patient flow performance. LOS represents the average duration of hospitalization among admitted patients. BOR represents the proportion of available hospital beds occupied during the relevant reporting period. BTR represents the frequency with which hospital beds were used by different patients within the reporting period. TOI represents the average duration for which a bed remained unoccupied between successive patients.

Clinical indicators represent aggregated monthly measures of disease burden and clinical severity among hospitalized patients. The complications included heart failure, cardiogenic shock, and DVT. Because the data were aggregated, these variables were treated as monthly indicators of institutional case mix rather than patient-specific predictors.

NDR was treated as a historical mortality-derived contextual indicator. To reduce target leakage, the NDR value from the preceding month, denoted as t−1t-1t−1, was used to predict GDR in the current month and was not calculated from deaths occurring during the same month as the target GDR. Because both NDR and GDR were derived from mortality data, the contribution of lagged NDR was interpreted separately from that of the clinical and operational indicators.

Data Preprocessing

All preprocessing procedures were implemented in the training portion of each cross-validation

fold. Imputation, categorical variable encoding, feature scaling, feature selection, and hyperparameter tuning were not performed on the complete dataset before validation. This pipeline-based procedure was used to reduce information leakage and overly optimistic estimates of the predictive performance.

Missing numerical values were replaced with median imputation. The median was calculated from the training data within each validation fold and was subsequently applied to the corresponding validation data. Categorical variables were transformed into numerical representations using label-encoding. Feature scaling was performed using the MinMaxScaler, which transformed the predictor values into a common numerical range. Scaling is particularly relevant to the multilayer perceptron because its gradient-based optimization procedure is sensitive to differences in feature magnitude.

Feature Selection and Feature Engineering

Feature selection was conducted cautiously because the dataset contained 36 monthly observations and 18 potential predictors. Recursive feature elimination and SHAP-based feature importance analyses were used as exploratory components of the modeling workflow. Recursive feature elimination iteratively removed less informative predictors based on model-based importance. SHAP values were used to quantify the contribution of each predictor to the model's output. The feature-selection results were interpreted as indicators of predictive relevance within the available dataset rather than as evidence of causal importance.

Feature engineering was performed to examine the potential interactions among the hospital operational indicators. Two prespecified interaction terms were constructed: LOS × BOR and BOR × BTR. These terms represent the possibility that combinations of hospital occupancy, patient turnover, and hospitalization duration contain predictive information beyond the corresponding individual variables. The interaction terms were selected using domain knowledge and exploratory correlation analyses. Given the limited number of observations, their contributions were interpreted as exploratory predictive signals rather than confirmed mechanisms of action.

Model Development

Linear regression was fitted to the conventional baseline model. It provided an interpretable reference against which the predictive performance of the machine learning models was compared.

The random forest model combined the predictions from multiple decision trees fitted to resampled training data and randomly selected predictor subsets. The XGBoost model sequentially fits boosted decision trees to reduce prediction errors from preceding iterations. The neural network was implemented as a feedforward multilayer perceptron capable of modeling nonlinear relationships and interactions among predictors.

A stacked ensemble model was developed by combining the predictions from the heterogeneous base learners. Random forest and XGBoost were used as base models, and their outputs were provided to a multilayer perceptron meta-learner to generate the final prediction. The stacking procedure was implemented within the validation workflow to prevent the meta-learner from being trained using the predictions derived from the validation observations. Because the number of monthly observations was limited, the ensemble analysis was considered to be exploratory.

Hyperparameter Optimization

The hyperparameters were optimized using a restricted grid search procedure embedded within the time-aware cross-validation framework.

The search space was deliberately limited to reduce the risk of overfitting due to the small sample size. For the random forest, the tuned hyperparameters included the number of trees and maximum tree depth. For XGBoost, the tuned hyperparameters included the learning rate and the number of boosting iterations. For the multilayer perceptron, the tuned hyperparameters included the hidden layer architecture and learning rate. Model selection emphasized internally validated predictive performance, stability across validation folds, and interpretability, rather than readiness for clinical or managerial deployment.

Validation Strategy

A five-fold time-aware cross-validation procedure was used because the observations represented chronological monthly sequences. Temporal ordering was preserved so that observations from later months were not used to predict the outcomes from earlier months. In

each fold, the models were trained using earlier observations and evaluated using subsequent observations according to a blocked or forward chaining structure.

All preprocessing, feature selection, feature engineering, and hyperparameter tuning procedures were conducted using only the training portion of each fold. The fitted procedures were then applied to the corresponding validation sections. The performance estimates were summarized across the validation folds. Because each fold contained relatively few observations, the resulting estimates were interpreted as exploratory measures of internal predictive performance rather than as evidence of external validity or generalizability.

Model Performance Evaluation

The model performance was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). The R^2 value describes the proportion of variation in the monthly GDR accounted for by the model. MAE represents the average absolute difference between the predicted and observed GDR values. MSE represents the average squared prediction error and assigns greater weight to larger errors. RMSE was calculated as the square root of the MSE and expressed in the same unit as the GDR.

Because the GDR was reported as deaths per 1,000 admissions, the MAE and RMSE were interpreted in the same unit. For example, an RMSE of 4.49 represents an average prediction error of approximately 4.49 deaths per 1,000 admissions and not 4.49 percentage points.

The primary model comparison emphasized R^2 and RMSE because the analysis sought to account for monthly variation in institutional GDR while limiting relatively large prediction errors. MAE was used as a complementary measure of the average absolute prediction error, and MSE was reported to describe the squared error. When the performance measures produced different model rankings, the differences were interpreted as trade-offs among the explained variation, average prediction error, and sensitivity to larger errors. The final model assessment considered the mean validation performance, consistency across folds, and model stability.

Result and Discussion

This study evaluated machine learning models for predicting monthly institutional GDR using aggregated clinical severity indicators and hospital operational performance metrics. The results include a model performance comparison, feature importance analysis, correlation analysis, distributional evidence, and regression diagnostics. All findings were interpreted at the institutional-month level.

Predictive Performance of Models

Predictive accuracy was measured using five-fold time-aware cross-validation with R^2, RMSE, and MAE. RMSE and MAE were expressed as deaths per 1,000 admissions, consistent with the unit used for the GDR in Table 1 and the figures. Higher R^2 and lower RMSE/MAE values indicate better predictive performance. The results are presented in Table 2.

Table 2. Cross-validation results of the predictive models. RMSE and MAE are expressed as deaths per 1,000 admissions

Model	Coefficient of Determination (R^2)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
Random Forest (RF)	0.733	5.83	3.54
Linear Regression (LR)	0.777	5.33	3.58
XGBoost (XGB)	0.575	7.36	4.59
Neural Network (NN)	0.801	5.03	2.74
Stacked Ensemble (NN+RF+XGB)	0.841	4.49	3.39

The stacked ensemble achieved the highest R^2 (0.841) and the lowest RMSE (4.49 deaths per 1,000 admissions), whereas the neural network achieved the lowest MAE (2.74 deaths per 1,000 admissions). Therefore, the ensemble was not superior across all metrics; rather, it showed the strongest performance for the explained variance and squared-error reduction, whereas the neural

network showed the smallest average absolute error. Similar recent mortality prediction studies have shown that nonlinear and ensemble-based models may improve predictive performance in hospital populations, but most of those studies used patient-level ICU or disease-specific data rather than aggregated monthly institutional indicators (He & Qiu, 2024; Jawadi et al., 2024; Z. Wu et al., 2025).

Recent reporting guidance for AI-based prediction models emphasizes that internally validated performance metrics should be interpreted together with sample size, uncertainty, calibration, and risk of bias. Therefore, the present R^2, RMSE, and MAE values should be interpreted as exploratory internal performance estimates rather than evidence of deployment readiness (Collins et al., 2024; Moons et al., 2025).

Figure 1 compares the internally cross-validated performance of the five predictive models across R^2, RMSE, and the MAE. The stacked ensemble achieved the highest R^2 (0.841) and the lowest RMSE (4.49 deaths per 1,000 admissions), whereas the neural network achieved the lowest MAE (2.74 deaths per 1,000 admissions). These findings indicate a trade-off among the explained variance, sensitivity to larger prediction errors, and average absolute error. The stacked ensemble was selected as the primary model because the prespecified model-selection rule prioritized R^2 and RMSE for monthly institutional monitoring. These findings support the cautious interpretation of model performance rather than an unconditional claim of ensemble superiority. However, the results should be interpreted as exploratory internal validation findings and require larger, multicenter, and prospective validation before implementation in hospital decision-support workflows.

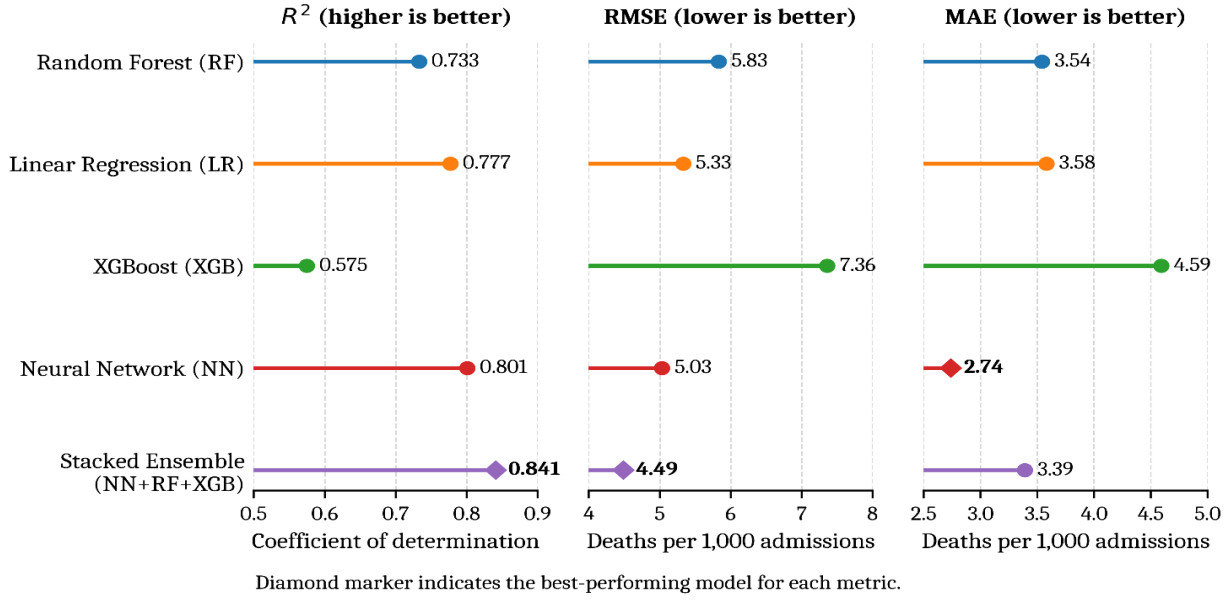

Figure 1. Cross-validated performance of five machine learning models for predicting monthly Gross Death Rate (GDR)

Although the stacked ensemble achieved the highest R^2 and lowest RMSE, the neural network produced the lowest MAE. This finding indicates a trade-off between the variance explanation, sensitivity to larger errors, and average absolute error. The stacked ensemble was selected as the primary model because the prespecified model selection rule prioritized R^2 and RMSE for monthly institutional monitoring.

The relatively weaker performance of XGBoost may reflect the small dataset size and aggregated institutional structure, which can limit the ability of boosting algorithms to learn stable patterns without overfitting the data. This interpretation is consistent with recent methodological work showing that clinical prediction models can be sensitive to data shifts, small validation samples, and changes in clinical workflow over time (Cabanillas Silva et al., 2024; Collins et al., 2024).

Accordingly, the model comparison results are best interpreted as internally generated

signals that require larger, multicenter, and prospective validation before being used in hospital decision-support workflows (Cabanillas Silva et al., 2024; Moons et al., 2025).

Feature Importance and SHAP Interpretation

Feature importance analysis was performed using the Random Forest model and supplemented with correlation analysis to identify predictors associated with the monthly GDR. SHAP-based interpretation is increasingly recommended in recent explainable ML mortality studies because it can help identify plausible predictors and improve the transparency of complex models (He & Qiu, 2024; Z. Wu et al., 2025).

Table 3 lists the five most influential predictors. The predictors were classified into three groups: mortality-derived historical indicators, aggregated clinical severity indicators, and operational performance indicators.

Table 3. Most powerful predictors of Gross Death Rate (GDR).

Feature	Importance Score
Net Death Rate	0.39
Heart Failure	0.35
Cardiogenic Shock	0.28
Length of Stay (LOS)	0.26
Bed Turnover Rate (BTR)	0.22

To reduce target leakage, the NDR was specified as a lagged historical predictor from the preceding month and did not include contemporaneous mortality information from the GDR prediction month. Nevertheless, because NDR and GDR are conceptually related mortality indicators, future sensitivity analyses should evaluate the model performance after excluding NDR and other mortality-derived predictors.

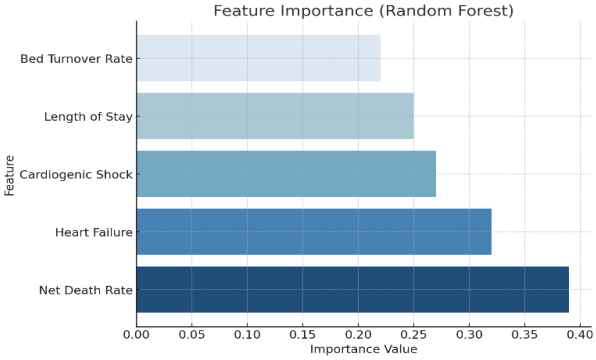

Figure 2. Feature Importance ranking from the Random Forest model.

Figure 2 supports the interpretation that historical mortality patterns, clinical severity indicators, and selected operational metrics contributed to GDR prediction. This interpretation is consistent with recent explainable mortality-prediction studies in which SHAP was used to clarify the contribution of clinical variables; however, the present predictors are aggregated monthly indicators and should not be interpreted as individual patient-level risk factors (He & Qiu, 2024; Z. Wu et al., 2025).

Correlation Analysis

To further clarify the inter-variable relationships in the dataset, a correlation heatmap (Figure 3) was created. The heatmap provides a visualization of the magnitude and direction of the correlation, which can aid in identifying patterns, and the correlation matrix shows the exact numerical relationship between the predictors.

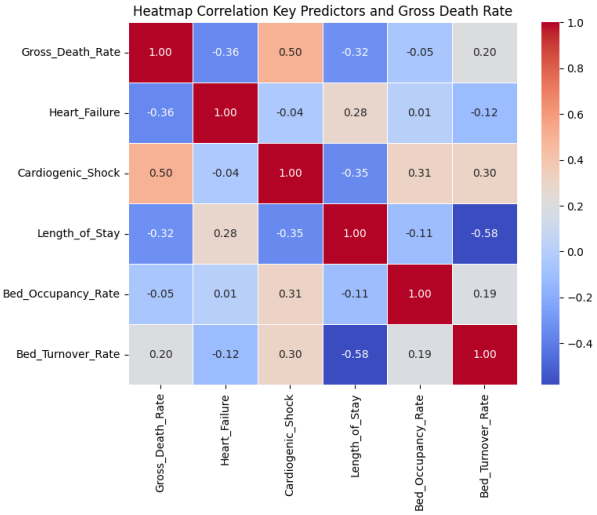

Figure 3. Correlation heatmap

Cardiogenic shock showed a moderate positive correlation with GDR ($r = 0.50$), suggesting that months with a higher cardiogenic shock burden tended to have a higher institutional mortality burden. This finding is clinically plausible because cardiogenic shock represents severe acute illness; however, the association remains descriptive because it was estimated from aggregated monthly data.

Because all variables were measured at the institutional-month level, the correlations shown in Figure 3 should be interpreted as

ecological associations. Recent hospital-performance literature shows that operational indicators, such as mortality rate, average length of stay, and bed occupancy rate, are important for institutional evaluation; however, their relationship with quality outcomes is context-dependent and potentially confounded by hospital case mix, demand, and service capacity (Bosque-Mercader & Siciliani, 2023; Talebpour et al., 2025).

Heart failure showed a moderate negative correlation with the gross death rate (GDR) (r = −0.36), while length of stay (LOS) was negatively correlated with GDR (r = −0.32). These associations should be interpreted descriptively rather than causally. The inverse correlation between heart failure and GDR may reflect variations in the monthly case mix, referral patterns, diagnostic coding, or other unmeasured institutional factors and should not be considered evidence of a protective effect. Likewise, the negative association between LOS and GDR does not indicate that longer hospitalization reduces mortality because aggregated ecological correlations may be influenced by discharge practices, patient complexity, referral flow, and other operational characteristics. Previous health services research has suggested that LOS may partially explain the relationship between bed occupancy and hospital quality, further emphasizing the need for caution when interpreting hospital-level operational indicators (Bosque-Mercader & Siciliani, 2023). Therefore, correlation analysis was used only to provide a descriptive context for the machine learning findings and not as evidence of causality or managerial effectiveness. Although hospital-level indicators may be useful for institutional surveillance, they cannot establish patient-level clinical effects without validation using individual-level, multicenter, and prospective data sets.

Regression Diagnostics and Distributional Evidence

To further explore the relationship between clinical severity indicators and institutional mortality burden, we examined the distribution of the monthly Gross Death Rate (GDR) across periods characterized by different levels of clinical severity indicators. Because the dataset consists of aggregated institutional observations rather than individual patient records, the analysis compared months with relatively higher and lower prevalence of specific clinical conditions. This approach enables a descriptive assessment of how fluctuations in clinical severity indicators correspond to variations in institutional-level mortality patterns.

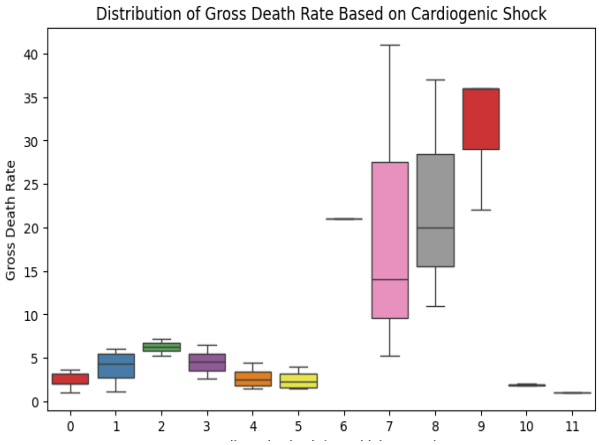

Figure 4. Distribution of monthly GDR according to aggregated monthly cardiogenic shock values

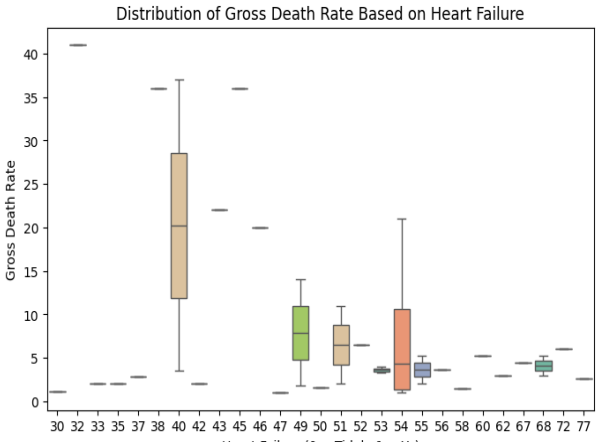

Figure 5. Distribution of monthly GDR according to aggregated monthly heart failure values

The distributional plots suggest that months with higher cardiogenic shock values tended to show higher GDR, whereas the relationship between heart failure values and GDR was more variable. These patterns were descriptive and should not be interpreted as causal or patient-level effects.

Ordinary Least Squares (OLS) regression was used as a supplementary diagnostic comparison. The residual plot (Figure 6) shows non-random patterns, suggesting that a simple linear specification may not fully capture the relationships among aggregated predictors and monthly GDR.

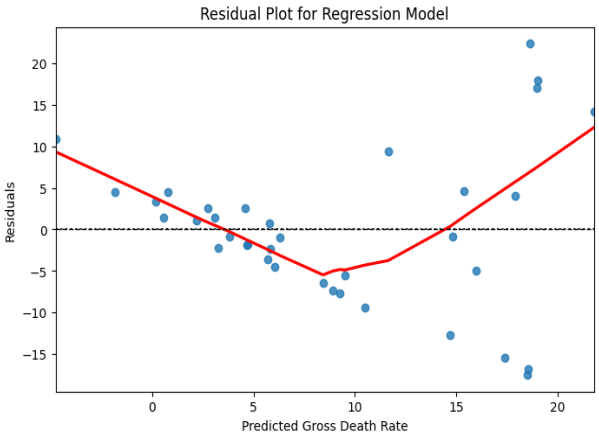

Figure 6. OLS residual plot

This diagnostic analysis supports the exploration of nonlinear machine learning approaches, but does not prove model superiority or clinical usefulness. Recent work on clinical-risk model shifts further indicates that model performance may change as patient populations, coding practices, and workflows evolve; therefore, temporal monitoring and external validation are necessary before routine use (Cabanillas Silva et al., 2024).

Principal Findings

This study investigated explainable machine learning models for predicting monthly institutional GDR using aggregated clinical severity indicators and hospital operational metrics. Three principal findings were obtained. First, the stacked ensemble achieved the highest R^2 and lowest RMSE, although the neural network had the lowest MAE. Second, lagged NDR was the strongest mortality-derived predictor, whereas heart failure and cardiogenic shock were important clinical severity indicators. Third, operational indicators contributed less directly but provided contextual information through the interaction terms. These findings should be compared cautiously with recent patient-level mortality prediction studies because those studies typically use individual ICU or disease-specific records rather than aggregated institutional-month observations (He & Qiu, 2024; Z. Wu et al., 2025).

Unlike many mortality prediction studies that evaluate patient-level risk using ICU or disease-specific data, this study focused on the institutional-level mortality burden measured by the GDR. Therefore, direct comparison with patient-level models is limited. The present findings are most relevant to hospital quality monitoring and health services research, where operational indicators are commonly used to assess institutional performance (Talebpour et al., 2025).

Ensemble Learning Superiority for Institutional-Level Prediction of Mortality

The stacked ensemble model achieved the highest explained variance and lowest RMSE, which is consistent with the theoretical advantage of combining heterogeneous learners and recent clinical ML studies showing that nonlinear models can capture complex risk patterns. However, because the dataset included only 36 monthly observations, this finding should be interpreted as preliminary internal evidence rather than proof of generalizable superiority (Collins et al., 2024; Moons et al., 2025).

These results do not eliminate the risk of overfitting. Complex models, such as neural networks, XGBoost, and stacked ensembles, may fit unstable patterns when the number of observations is small. This concern is consistent with PROBAST+AI, which emphasizes risk-of-bias assessment, applicability, and validation of prediction models using artificial intelligence methods (Moons et al., 2025).

Therefore, the observed performance gain should be described as an internally validated pattern rather than as evidence that the model is ready for clinical or managerial implementation.

Greater Predictive Contribution of Clinical and Historical Mortality Indicators Compared with Operational Indicators

Feature importance, correlation, and diagnostic analyses suggested that lagged mortality history and aggregated clinical severity indicators had stronger direct contributions to GDR prediction than individual operational metrics. This finding aligns with recent explainable ML mortality studies showing that clinical severity markers often dominate predictive models, while SHAP or feature-importance methods help clarify how predictors contribute to model outputs (He & Qiu, 2024; Z. Wu et al., 2025).

Operational metrics such as LOS, BOR, and BTR had modest direct predictive importance in this dataset. Nevertheless, hospital-performance research supports their inclusion in institutional monitoring because the average length of stay, mortality rate, and bed occupancy rate are

among the commonly used hospital key performance indicators (Talebpour et al., 2025). Evidence from the English National Health Service also suggests that bed occupancy may be associated with hospital quality outcomes, although such associations are context-dependent and partly explained by LOS (Bosque-Mercader & Siciliani, 2023).

Contribution of Feature Engineering and Interaction Effect

Feature importance and correlation analyses suggested that mortality-derived historical indicators and clinical severity indicators were important contributors to the monthly GDR prediction. Cardiogenic shock was positively associated with the GDR, consistent with its role as a marker of severe clinical acuity. The negative association between heart failure and GDR was interpreted cautiously as an ecological pattern that may reflect case mix or referral dynamics rather than a causal relationship. This cautious interpretation is consistent with the guidance that prediction model findings should not be converted into causal claims without an appropriate study design and validation (Collins et al., 2024; Moons et al., 2025).

Operational metrics, such as LOS, BOR, and BTR, showed modest direct importance. Their role may be contextual rather than independently causal, helping the model to represent hospital flow and capacity conditions that coincide with mortality variation. Recent KPI literature supports the role of these indicators in hospital evaluation but also reinforces that they should be interpreted together with outcome, safety, and service quality measures (Talebpour et al., 2025).

This supports a balanced interpretation: operational indicators add institutional context to clinical and historical mortality signals; however, the present analysis does not establish them as direct causal determinants of mortality. This interpretation follows the recent guidance for prediction-model discussions, which recommends separating predictive associations from causal or implementation claims (Collins et al., 2024; Moons et al., 2025).

Engineered interaction features may have helped the models capture the relationships among operational indicators, such as the combined influence of bed use and patient flow. However, these interaction effects should be interpreted cautiously because the small sample size limits the stability of the relationships between complex features. Further studies using longer time series and multicenter data are needed to determine whether these interactions are reproducible across hospital settings (Cabanillas Silva et al., 2024; Moons et al., 2025).

These results support a hybrid modeling perspective in which operational indicators complement clinical and mortality indicators. They should not be interpreted as evidence that administrative indicators are direct levers for reducing mortality but as contextual signals for institutional monitoring, in line with recent hospital KPI literature (Bosque-Mercader & Siciliani, 2023; Talebpour et al., 2025).

Clinical, Managerial, and Methodological Implications

From a clinical and managerial perspective, these findings may inform the future development of institutional mortality monitoring tools. However, the model is not ready for direct use in bed management, escalation protocols, or operational workflows without external validation, calibration assessment, usability assessment, or prospective impact evaluation. This position is consistent with TRIPOD+AI and PROBAST+AI, which emphasize transparent reporting, risk-of-bias assessment, and validation before AI-based prediction models are used in practice (Collins et al., 2024; Moons et al., 2025).

For hospital management, the findings suggest that operational indicators should be interpreted in conjunction with clinical severity and historical mortality patterns. Administrative efficiency measures should not be treated as direct causal determinants of mortality based on this exploratory aggregated analysis. Instead, they should be used as contextual dashboard indicators, consistent with the recent hospital KPI literature (Bosque-Mercader & Siciliani, 2023; Talebpour et al., 2025).

Methodologically, this study illustrates how machine learning and feature importance analysis can be applied to institutional-level hospital indicators. Nevertheless, transparent reporting, leakage control, uncertainty estimation, temporal monitoring, and validation using larger datasets are required to strengthen future studies. Recent evidence on longitudinal model shifts demonstrates that clinical prediction models may require ongoing performance monitoring when deployed across

hospitals or over time (Cabanillas Silva et al., 2024).

Limitations

This study had several limitations. The analysis included only 36 monthly observations from a single referral hospital, which limited the statistical power and increased the risk of overfitting. Because all variables were aggregated at the institutional-month level, the findings cannot be interpreted as patient-level risk estimates. The inclusion of lagged NDR as a mortality-related predictor also requires a sensitivity analysis excluding mortality-derived variables.

In addition, the models were not externally or prospectively validated, and their calibration and workflow implementation were not assessed. Therefore, the findings should be considered preliminary and interpreted in accordance with the TRIPOD+AI and PROBAST+AI recommendations. GDR is reported as deaths per 1,000 admissions.

Conclusion

This exploratory single-center study developed and internally evaluated explainable machine learning models for predicting the monthly institutional gross death rate using aggregated clinical severity and hospital operational indicators. The stacked ensemble showed the strongest overall predictive performance, whereas the neural network produced the lowest average absolute error.

Model interpretation identified the lagged net death rate as the most influential predictor, with heart failure and cardiogenic shock contributing substantially. Operational indicators provide additional contextual information but have smaller direct contributions. These findings support the feasibility of XGBoost for hospital-level mortality surveillance; however, they should not be interpreted as patient-level risk estimates or used for clinical or managerial decisions. External multicenter validation, calibration assessment, sensitivity analyses excluding mortality-derived predictors, and prospective workflow evaluation are required before the implementation of this model.

The models may be used as research prototypes to guide the development of hospital-level mortality surveillance systems but not as standalone tools for clinical or administrative decision-making. Future studies should validate the models using larger multicenter datasets, assess their calibration and temporal performance, perform sensitivity analyses excluding mortality-derived predictors, and prospectively evaluate their integration into hospital monitoring workflows.

References

Bosque-Mercader, L., & Siciliani, L. (2023). The association between bed occupancy rates and hospital quality in the English National Health Service. *The European Journal of Health Economics, 24*(2), 209–236. https://doi.org/10.1007/s10198-022-01464-8

Cabanillas Silva, P., Sun, H., Rezk, M., Roccaro-Waldmeyer, D. M., Fliegenschmidt, J., Hulde, N., von Dossow, V., Meesseman, L., Depraetere, K., Stieg, J., Szymanowsky, R., & Dahlweid, F.-M. (2024). Longitudinal model shifts of machine learning–based clinical risk prediction models: Evaluation study of multiple use cases across different hospitals. *Journal of Medical Internet Research, 26*, e51409. https://doi.org/10.2196/51409

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., . . . Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Deng, T., Hamdan, H., Yaakob, R., & Kasmiran, K. A. (2023). Personalized federated learning for in-hospital mortality prediction of multi-center ICU. *IEEE Access, 11*, 11652–11663. https://doi.org/10.1109/ACCESS.2023.3241488

Deschepper, M., Waegeman, W., Vogelaers, D., & Eeckloo, K. (2020). Using structured pathology data to predict hospital-wide mortality at admission. *PLOS ONE, 15*(6),

e0235117. https://doi.org/10.1371/journal.pone.0235117

Estiri, H., Strasser, Z. H., Klann, J. G., Naseri, P., Wagholikar, K. B., & Murphy, S. N. (2021). Predicting COVID-19 mortality with electronic medical records. *npj Digital Medicine, 4*(1), Article 15. https://doi.org/10.1038/s41746-021-00383-x

Fang, C., Pan, Y., Zhao, L., Niu, Z., Guo, Q., & Zhao, B. (2022). A machine learning–based approach to predict prognosis and length of hospital stay in adults and children with traumatic brain injury: Retrospective cohort study. *Journal of Medical Internet Research, 24*(12), e41819. https://doi.org/10.2196/41819

Giwangkancana, G. W., Anina, H. N., & Sukandar, H. (2024). Predicting end-of-life in a hospital setting. *Journal of Multidisciplinary Healthcare, 17*, 619–627. https://doi.org/10.2147/JMDH.S443425

Halasz, G., Sperti, M., Villani, M., Michelucci, U., Agostoni, P., Biagi, A., Rossi, L., Botti, A., Mari, C., Maccarini, M., Pura, F., Roveda, L., Nardecchia, A., Mottola, E., Nolli, M., Salvioni, E., Mapelli, M., Deriu, M. A., Piga, D., & Piepoli, M. (2021). A machine learning approach for mortality prediction in COVID-19 pneumonia: Development and evaluation of the Piacenza score. *Journal of Medical Internet Research, 23*(5), e29058. https://doi.org/10.2196/29058

He, B., & Qiu, Z. (2024). Development and validation of an interpretable machine learning for mortality prediction in patients with sepsis. *Frontiers in Artificial Intelligence, 7*, 1348907. https://doi.org/10.3389/frai.2024.1348907

Hsu, C.-N., Liu, C.-L., Tain, Y.-L., Kuo, C.-Y., & Lin, Y.-C. (2020). Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: Development and validation study. *Journal of Medical Internet Research, 22*(8), e16903. https://doi.org/10.2196/16903

Huang, T., Le, D., Yuan, L., Xu, S., & Peng, X. (2023). Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit. *PLOS ONE, 18*(1), e0280606. https://doi.org/10.1371/journal.pone.0280606

Jawadi, Z., He, R., Srivastava, P. K., Fonarow, G. C., Khalil, S. O., Krishnan, S., Eskin, E., Chiang, J. N., & Nsair, A. (2024). Predicting in-hospital mortality among patients admitted with a diagnosis of heart failure: A machine learning approach. *ESC Heart Failure, 11*(5), 2490–2498. https://doi.org/10.1002/ehf2.14796

König, S., Pellissier, V., Hohenstein, S., Leiner, J., Meier-Hellmann, A., Kuhlen, R., Hindricks, G., & Bollmann, A. (2022). From population- to patient-based prediction of in-hospital mortality in heart failure using machine learning. *European Heart Journal – Digital Health, 3*(2), 307–310. https://doi.org/10.1093/ehjdh/ztac012

Kwun, J. S., Ahn, H. B., Kang, S. H., Yoo, S., Kim, S., Song, W., Hyun, J., Oh, J. S., Baek, G., & Suh, J. W. (2025). Developing a machine learning model for predicting 30-day major adverse cardiac and cerebrovascular events in patients undergoing noncardiac surgery: Retrospective study. *Journal of Medical Internet Research, 27*, e66366. https://doi.org/10.2196/66366

Lee, B., Kim, K., Hwang, H., Kim, Y. S., Chung, E. H., Yoon, J. S., Cho, H. J., & Park, J. D. (2021). Development of a machine learning model for predicting pediatric mortality in the early stages of intensive care unit admission. *Scientific Reports, 11*(1), Article 1263. https://doi.org/10.1038/s41598-020-80474-z

Lee, S. W., Lee, H. C., Suh, J., Lee, K. H., Lee, H., Seo, S., Kim, T. K., Lee, S. W., & Kim, Y. J. (2022). Multi-center validation of machine learning model for preoperative prediction of postoperative mortality. *npj Digital Medicine, 5*(1), Article 91. https://doi.org/10.1038/s41746-022-00625-6

Li, C., Zhang, Z., Ren, Y., Nie, H., Lei, Y., Qiu, H., Xu, Z., & Pu, X. (2021). Machine learning based early mortality prediction in the emergency department. *International Journal of Medical Informatics, 155*, 104570. https://doi.org/10.1016/j.ijmedinf.2021.104570

Li, L., Ding, L., Zhang, Z., Zhou, L., Zhang, Z., Xiong,

Y., Hu, Z., & Yao, Y. (2023). Development and validation of machine learning–based models to predict in-hospital mortality in life-threatening ventricular arrhythmias: Retrospective cohort study. *Journal of Medical Internet Research, 25*, e47664. https://doi.org/10.2196/47664

Li, M., Han, S., Liang, F., Hu, C., Zhang, B., Hou, Q., & Zhao, S. (2024). Machine learning for predicting risk and prognosis of acute kidney disease in critically ill elderly patients during hospitalization: Internet-based and interpretable model study. *Journal of Medical Internet Research, 26*, e51354. https://doi.org/10.2196/51354

Lv, H., Yang, X., Wang, B., Wang, S., Du, X., Tan, Q., Hao, Z., Liu, Y., Yan, J., & Xia, Y. (2021). Machine learning–driven models to predict prognostic outcomes in patients hospitalized with heart failure using electronic health records: Retrospective study. *Journal of Medical Internet Research, 23*(4), e24996. https://doi.org/10.2196/24996

Maheswari, B. U., Ashik, F., George, A., & Jose, A. (2023). In-hospital mortality prognosis: Unmasking patterns using data science and explainable AI. In *2023 9th International Conference on Signal Processing and Communication (ICSC)* (pp. 356–361). IEEE. https://doi.org/10.1109/ICSC60394.2023.10441356

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Liu, N., . . . Van Smeden, M. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388*, e082505. https://doi.org/10.1136/bmj-2024-082505

Muralitharan, S., Nelson, W., Di, S., McGillion, M., Devereaux, P. J., Barr, N. G., & Petch, J. (2021). Machine learning–based early warning systems for clinical deterioration: Systematic scoping review. *Journal of Medical Internet Research, 23*(2), e25187. https://doi.org/10.2196/25187

Naemi, A., Schmidt, T., Mansourvar, M., Naghavi-Behzad, M., Ebrahimi, A., & Wiil, U. K. (2021). Machine learning techniques for mortality prediction in emergency departments: A systematic review. *BMJ Open, 11*(11), e052663. https://doi.org/10.1136/bmjopen-2021-052663

Nie, X., Cai, Y., Liu, J., Liu, X., Zhao, J., Yang, Z., Wen, M., & Liu, L. (2021). Mortality prediction in cerebral hemorrhage patients using machine learning algorithms in intensive care units. *Frontiers in Neurology, 11*, 610531. https://doi.org/10.3389/fneur.2020.610531

Qiao, E. M., Qian, A. S., Nalawade, V., Voora, R. S., Kotha, N. V., Vitzthum, L. K., & Murphy, J. D. (2022). Evaluating high-dimensional machine learning models to predict hospital mortality among older patients with cancer. *JCO Clinical Cancer Informatics, 6*, e2100186. https://doi.org/10.1200/CCI.21.00186

Seki, T., Kawazoe, Y., & Ohe, K. (2021). Machine learning–based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data. *PLOS ONE, 16*(2), e0246640. https://doi.org/10.1371/journal.pone.0246640

Shah, N., Konchak, C., Chertok, D., Au, L., Kozlov, A., Ravichandran, U., McNulty, P., Liao, L., Steele, K., Kharasch, M., Boyle, C., Hensing, T., Lovinger, D., Birnberg, J., Solomonides, A., & Halasyamani, L. (2020). Clinical Analytics Prediction Engine (CAPE): Development, electronic health record integration and prospective validation of hospital mortality, 180-day mortality and 30-day readmission risk prediction models. *PLOS ONE, 15*(8), e0238065. https://doi.org/10.1371/journal.pone.0238065

Sinha, S., Dong, T., Dimagli, A., Judge, A., & Angelini, G. D. (2024). A machine learning algorithm–based risk prediction score for in-hospital/30-day mortality after adult cardiac surgery. *European Journal of Cardio-Thoracic Surgery, 66*(4), ezae368. https://doi.org/10.1093/ejcts/ezae368

Talebpour, A., Sadeghi-Bazargani, H., Janati, A.,

Pashazadeh, F., & Gholizadeh, M. (2025). Crucial key performance indicators for hospital evaluation: A scoping review. *Journal of Education and Health Promotion, 14*(1), Article 195. https://doi.org/10.4103/jehp.jehp_2102_23

Thorsen-Meyer, H.-C., Nielsen, A. B., Nielsen, A. P., Kaas-Hansen, B. S., Toft, P., Schierbeck, J., Strøm, T., Chmura, P. J., Heimann, M., Dybdahl, L., Spangsege, L., Hulsen, P., Belling, K., Brunak, S., & Perner, A. (2020). Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health, 2*(4), e179–e191. https://doi.org/10.1016/S2589-7500(20)30018-2

Trentino, K. M., Schwarzbauer, K., Mitterecker, A., Hofmann, A., Lloyd, A., Leahy, M. F., Tschoellitsch, T., Böck, C., Hochreiter, S., & Meier, J. (2022). Machine learning–based mortality prediction of patients at risk during hospital admission. *Journal of Patient Safety, 18*(5), 494–498. https://doi.org/10.1097/PTS.0000000000000957

Warman, P. I., Seas, A., Satyadev, N., Adil, S. M., Kolls, B. J., Haglund, M. M., Dunn, T. W., & Fuller, A. T. (2022). Machine learning for predicting in-hospital mortality after traumatic brain injury in both high-income and low- and middle-income countries. *Neurosurgery, 90*(5), 605–612. https://doi.org/10.1227/neu.0000000000001898

Wu, M., & Gao, H. (2023). A prediction model for in-hospital mortality in intensive care unit patients with metastatic cancer. *Frontiers in Surgery, 10*, 992936. https://doi.org/10.3389/fsurg.2023.992936

Wu, Z., Li, M., Xu, Z., & Liu, G. (2025). Machine learning model development and validation using SHAP: Predicting 28-day mortality risk in pulmonary fibrosis patients. *BMC Medical Informatics and Decision Making, 25*(1), Article 382. https://doi.org/10.1186/s12911-025-03172-8

Yun, K., Oh, J., Hong, T. H., & Kim, E. Y. (2021). Prediction of mortality in surgical intensive care unit patients using machine learning algorithms. *Frontiers in Medicine, 8*, 621861. https://doi.org/10.3389/fmed.2021.621861