UDC 004.8
IRSTI 28.23.34

https://doi.org/10.55452/1998-6688-2026-23-3-204-217

[1]*Kunichik V.G.,
PhD student, ORCID ID: 0009-0004-8580-7731,
*e-mail: kunichikval@gmail.com
[1]Salykova O.S.,
Cand.Tech.Sc., Associate Professor,
ORCID ID: 0000-0002-8681-4552, e-mail: solga0603@mail.ru

[1]Akhmet Baitursynuly Kostanay Regional University, Kostanay, Kazakhstan

EARLY PREDICTION OF IN-HOSPITAL MORTALITY USING INTERPRETABLE AND CALIBRATED ENSEMBLE MACHINE LEARNING MODELS

Abstract

Early identification of patients at high risk of in-hospital mortality is a persistent challenge in clinical practice, particularly during infectious disease outbreaks such as COVID-19, when decisions must be made at the time of hospital admission using incomplete information. Although machine learning methods have been widely applied to mortality prediction, many existing models are developed and assessed under conditions that limit their usefulness for early triage, including insufficient attention to probability reliability, temporal generalization, and decision-oriented assessment. This work examines early-stage mortality risk prediction using routinely available clinical data collected at the time of hospital admission. An ensemble of gradient boosting models is developed and analyzed under strict temporal separation to reflect real-world deployment conditions. Particular emphasis is placed on the reliability of predicted probabilities, with post-hoc isotonic calibration applied to improve alignment between predicted risks and observed outcomes. Model performance is assessed using complementary discrimination and calibration metrics, while clinical relevance is examined through decision curve analysis across plausible operating thresholds. Evaluation on an independent, temporally held-out test set shows that ensemble aggregation improves the stability of risk estimates, while calibration yields more consistent threshold-based behavior without altering ranking performance. Decision curve analysis indicates that calibrated predictions provide higher net benefit than default decision strategies across a broad range of early triage thresholds. These findings highlight that, in early clinical decision-making, probability reliability plays a critical role alongside discrimination. The presented framework offers a methodologically robust approach to early in-hospital mortality risk assessment under realistic clinical and temporal constraints.

Keywords: machine learning, in-hospital mortality, early clinical triage, ensemble models, probability calibration, decision curve analysis, clinical decision support.

Received January 13, 2026; revised April 2, 2026; accepted May 20, 2026.

Introduction

Recent advances in information and communication technologies have enabled the widespread adoption of data-driven decision support systems in high-impact domains, including healthcare. The increasing availability of electronic health records and national clinical registries has stimulated the application of machine learning methods for predictive modeling, risk stratification, and outcome forecasting across a broad range of clinical tasks. In this context, increasing attention has been paid to early prediction of adverse clinical outcomes using routinely available clinical information collected at the initial stages of patient assessment, where decisions must often be made before detailed laboratory or imaging data become available [1–2].

Machine learning models have demonstrated promising results in mortality prediction tasks across different clinical settings, including infectious disease outbreaks such as COVID-19. Ensemble methods, tree-based algorithms, and neural networks are frequently reported to outperform traditional statistical approaches in terms of discriminative ability [3–5]. However, strong predictive discrimination alone does not guarantee reliable or clinically meaningful risk estimates. In early clinical decision-making settings, even well-discriminating models may produce risk estimates that are difficult to interpret or operationalize in practice when absolute probabilities are not sufficiently reliable [3].

Another critical challenge limiting the real-world deployment of machine learning models in healthcare is interpretability. Many high-performing models operate as "black boxes", providing limited insight into the mechanisms underlying their predictions. This lack of transparency reduces clinicians' trust and complicates the integration of such models into routine clinical workflows. As a result, recent research increasingly emphasizes interpretable machine learning approaches that enable assessment of feature contributions and facilitate understanding of individual risk predictions [6–8].

In addition to discrimination and interpretability, the reliability of predicted probabilities has emerged as a central methodological consideration for clinical prediction models. When predicted risks are intended to support patient-level decision-making, calibration becomes essential, particularly under temporal distribution shift and evolving clinical conditions [9]. Closely related to this issue is the evaluation of clinical utility. Traditional performance metrics, such as accuracy or the area under the receiver operating characteristic curve, do not directly reflect the consequences of model-based decisions. Decision curve analysis has therefore been proposed as an effective framework for assessing the net benefit of predictive models across a range of risk thresholds and for explicitly linking model outputs to actionable decision strategies [10–11].

Despite the growing body of literature on machine learning–based mortality prediction, many studies continue to focus primarily on improving predictive performance or proposing novel model architectures. Methodological aspects related to early-stage applicability, probability reliability, and the translation of model outputs into clinically meaningful decision thresholds remain insufficiently emphasized [12]. Furthermore, the use of lightweight clinical data available at the early stages of hospitalization remains underexplored, despite its importance for timely clinical triage and prospective decision support [13].

In this context, the aim of this study is to develop and evaluate an interpretable ensemble machine learning approach for early prediction of in-hospital mortality using routinely available admission-time clinical data collected at the initial stage of hospitalization. We hypothesize that combining ensemble learning with calibration and interpretability techniques can enhance the reliability and clinical usefulness of early mortality risk predictions when only limited clinical information is available. To investigate this hypothesis, large-scale registry data are analyzed under strict temporal separation, ensemble machine learning models are trained and post-hoc calibrated, discrimination and probability reliability are jointly evaluated, and potential clinical utility is assessed using decision curve analysis and operational risk thresholds. This study contributes a rigorously validated, calibration-aware ensemble framework for early in-hospital mortality prediction, explicitly linking probabilistic risk estimates to clinically actionable decision thresholds under temporal dataset shift. The novelty of this study lies in its calibration-aware, decision-oriented evaluation of early-stage ensemble models under strict temporal validation rather than in proposing a new algorithmic architecture.

Materials and methods

Data source and cohort construction

This study was based on anonymized secondary registry data of hospitalized patients with confirmed COVID-19 infection, with in-hospital mortality defined as a binary outcome. The registry represents a large-scale, real-world clinical data source reflecting routine hospital practice across

multiple pandemic waves. To support reproducibility, the data were consolidated into a unified year-indexed panel with a fixed feature schema across years and stored in Parquet format.

The final analytical cohort comprised approximately 2.53 million hospital admissions, with an overall in-hospital mortality proportion of 0.284. Notably, outcome prevalence varied substantially across calendar years, with mortality rates of 0.305 in 2020, 0.299 in 2021, and 0.214 in 2022, indicating pronounced temporal heterogeneity in baseline risk. This variability reflects changes in patient characteristics, clinical management, and pandemic dynamics over time and motivated the explicit consideration of temporal generalization in subsequent modeling stages.

At the cohort assembly stage, data processing decisions were deliberately conservative. Aggressive record exclusion, extensive feature engineering, or heavy imputation were intentionally avoided in order to preserve the natural structure of the registry data. Data quality was assessed through reproducible exploratory checks of demographic distributions, missingness patterns, and indicator correlations; these diagnostics confirmed expected clinical gradients and the absence of critical schema inconsistencies. This design choice was motivated by the study's focus on probability calibration and clinical utility, as excessive preprocessing may introduce implicit assumptions that distort risk distributions and bias downstream decision-curve analyses. By retaining the original data structure as much as possible, the constructed cohort provides a realistic basis for evaluating early-stage risk prediction under real-world conditions.

Validation design and leakage control (temporal generalization)

Consistent with the exploratory cohort-level analysis described above, all validation and evaluation procedures were explicitly designed to account for pronounced temporal non-stationarity and baseline risk shifts observed across pandemic waves. To approximate real-world deployment conditions and explicitly address dataset shift over time, a strict temporal validation design was adopted [14–16]. All models were trained using data from 2020–2021 and evaluated on an independent, chronologically held-out test set from 2022, ensuring that no information from the future period was available at any stage of model training, calibration, or model selection.

This validation strategy was chosen in response to substantial temporal drift observed in both outcome prevalence and age-stratified risk distributions across calendar years. Under such conditions, risk estimates, predicted probabilities, and clinically meaningful decision thresholds learned on earlier data cannot be assumed to remain valid in later periods without explicit re-validation [14, 16]. In particular, evaluation protocols based on random splits or cross-validation may yield overly optimistic performance estimates by implicitly sharing information across time and thereby introducing temporal leakage [15].

By enforcing a fully out-of-time evaluation, the adopted design provides a more realistic assessment of temporal generalization, calibration stability, and clinical utility under evolving epidemiological and clinical conditions [14, 16]. This approach is especially critical for early-stage mortality risk prediction, where models are intended to support prospective decision-making and must remain robust to non-stationary data-generating processes.

Feature set and preprocessing principles (early triage constraints)

The predictive pipeline was deliberately constrained to an early-stage (EHR-lite) feature set, consisting exclusively of variables routinely available at the time of initial hospital admission and suitable for early clinical triage. This restriction reflects the intended use case of the models, namely prospective risk stratification before the availability of laboratory dynamics, imaging results, or in-hospital treatment trajectories.

Preprocessing emphasized consistency and transparency rather than maximal feature enrichment [12]. All yearly data extracts were mapped to a unified feature schema to ensure comparability across pandemic waves, while inclusion criteria were applied uniformly to avoid introducing year-specific biases. Importantly, preprocessing decisions were explicitly separated from model training, allowing their effects on probability calibration and downstream threshold-based decision-making to be evaluated in isolation.

Operations such as missing-value imputation, construction of derived variables, and additional filtering were treated as methodological choices requiring explicit justification, rather than as default preprocessing steps. This design reduces the risk of inadvertently encoding outcome-related information into the feature set and supports interpretability of predicted risk distributions under early-triage constraints.

Model family and ensemble construction

Consistent with best practices for clinical risk prediction under dataset shift, multiple model families were evaluated within a unified training and validation framework to assess robustness of early mortality risk prediction [17]. Candidate models included regularized logistic regression (L2), elastic net logistic regression, random forest, histogram-based gradient boosting, LightGBM, and XGBoost. All models were trained using the same early-stage (EHR-lite) feature set and evaluated under an identical temporal validation protocol.

Model selection was guided not only by discriminative performance but also by stability of predicted probabilities and robustness across time. While individual models achieved comparable AUROC values, tree-based gradient boosting methods consistently demonstrated superior precision–recall characteristics and more stable probability distributions under temporal generalization, in line with prior observations in clinical machine learning.

Based on these observations, the final predictive model was constructed as an ensemble of three high-performing gradient boosting learners with complementary inductive biases. Ensemble aggregation was performed by averaging predicted probabilities across the selected base models. This simple aggregation strategy was chosen to reduce variance and mitigate model-specific error patterns while preserving interpretability and calibration properties. More complex stacking architectures were deliberately avoided to limit overfitting and to preserve interpretability of the modeling pipeline.

The resulting ensemble demonstrated improved stability of risk estimates compared to individual models and served as the primary prediction engine for subsequent calibration, discrimination, and decision-analytic evaluation.

Probability calibration

Because model outputs were intended to support threshold-based clinical decision-making, particular emphasis was placed on the reliability of predicted probabilities rather than ranking performance alone [9]. In early triage settings, inaccurate probability estimates directly affect threshold selection and may lead to inappropriate clinical actions, even when discrimination metrics appear satisfactory.

To address this, post-hoc probability calibration was applied using isotonic regression, a non-parametric calibration method that does not impose parametric assumptions on the relationship between raw model scores and observed outcome frequencies [18–19]. Calibration models were trained exclusively on the training data, ensuring that evaluation on the temporally held-out test set reflected true out-of-sample probability reliability without information leakage.

Both uncalibrated (raw) and calibrated probability estimates were retained for downstream analysis. This design enabled direct comparison of discrimination metrics, calibration quality, and decision-curve behavior before and after calibration. Importantly, isotonic calibration preserved ranking performance while improving alignment between predicted risks and observed outcome frequencies, thereby supporting more consistent and interpretable threshold-based decision-making.

Evaluation metrics (discrimination and reliability)

Model performance was evaluated on the independent 2022 temporally held-out test set using complementary metrics designed to capture both discriminative ability and probability reliability [20]. This dual evaluation strategy reflects the study's focus on clinically meaningful risk estimation rather than ranking performance alone.

Discrimination was assessed using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (PR-AUC), the latter being particularly informative given the non-negligible prevalence of in-hospital mortality. To quantify the accuracy

of probabilistic predictions, overall calibration quality was evaluated using the Brier score, which jointly reflects calibration and refinement.

For the final isotonic-calibrated ensemble, performance on the 2022 test set reached AUROC = 0.812, PR-AUC = 0.483, and Brier score = 0.140, indicating strong discrimination accompanied by reliable probability estimates. Reporting discrimination and reliability metrics jointly enables a more comprehensive assessment of model suitability for threshold-based clinical decision support.

Clinical utility via Decision Curve Analysis

To assess whether improvements in predictive performance translate into clinically meaningful decision support, we evaluated model utility using Decision Curve Analysis (DCA) on the independent 2022 test set [10–11]. DCA quantifies net benefit by comparing model-assisted decision-making with default reference strategies, specifically "treat all" and "treat none," across a range of clinically relevant probability thresholds.

The isotonic-calibrated ensemble demonstrated consistent dominance over both reference strategies across a broad threshold interval, approximately 0.15–0.50, indicating a favorable balance between correct identification of high-risk patients and avoidance of excessive false-positive decisions [21–22]. This pattern indicates that improved probability reliability supports more consistent threshold-based clinical decision-making rather than improvements in ranking performance alone.

In addition, an operational triage interval between 0.20 and 0.35 was identified as a practically relevant region in which the calibrated ensemble maintained positive net benefit while remaining robustly superior to baseline strategies. This interval represents a clinically plausible compromise between sensitivity and specificity under early triage constraints and illustrates how calibrated risk estimates can support informed threshold selection in real-world settings.

Implementation and reproducibility

All preprocessing, modeling, calibration, and evaluation steps were implemented within reproducible computational workflows, ensuring traceability of analytical decisions across the entire pipeline. Data handling, model training, validation, and performance assessment followed a fixed and auditable sequence designed to minimize procedural variability and support methodological transparency.

To facilitate reproducibility, non-sensitive artifacts–including the preprocessing schema, feature definitions, validation protocols, and evaluation procedures–can be made available upon reasonable request. Code components required to reproduce the analytical workflow and key experimental results may also be provided, subject to institutional and regulatory constraints.

Access to raw patient-level data is restricted due to privacy, ethical, and legal considerations. These data protection requirements are consistent with the use of secondary clinical registry data and do not affect the reproducibility of the methodological framework presented in this study.

Results and discussion

Cohort characteristics and descriptive analysis

The final analytical cohort comprised approximately 2.53 million hospital admissions with confirmed COVID-19 infection and a binary outcome defined as in-hospital mortality. The dataset represents a large-scale, real-world clinical registry covering multiple pandemic waves and reflecting routine hospital practice under varying epidemiological and organizational conditions.

As an initial descriptive and exploratory analysis step, cohort-level characteristics and outcome dynamics were examined to assess data integrity, temporal structure, and baseline risk heterogeneity prior to predictive modeling. This exploratory analysis was intended to verify clinically plausible gradients, identify temporal non-stationarity, and inform subsequent choices related to validation design, calibration strategy, and decision-oriented evaluation.

Overall in-hospital mortality across the full cohort was 0.284. Substantial temporal variation in outcome prevalence was observed across calendar years, with mortality rates of 0.305 in 2020, 0.299 in 2021, and 0.214 in 2022. This marked decline in observed mortality during later pandemic periods

reflects changes in patient composition, clinical management, and broader epidemiological dynamics over time.

All records were harmonized into a unified feature schema to ensure consistency across yearly data extracts. The cohort construction deliberately preserved the original structure of the registry without aggressive record exclusion or extensive feature transformation. As a result, the observed descriptive patterns reflect naturally occurring variability in routinely collected clinical data rather than artifacts introduced by preprocessing.

Baseline descriptive characteristics demonstrated pronounced heterogeneity in patient-level risk across the cohort. Temporal shifts in outcome prevalence translated into changes in baseline risk distributions, underscoring the non-stationary nature of the data-generating process. These findings motivated the use of strict chronological separation between training and evaluation data and the explicit assessment of model robustness under temporal distribution shift.

Figure 1 summarizes the exploratory temporal analysis of outcome dynamics by presenting the weekly in-hospital mortality rate over the full observation period together with leakage-safe smoothing based on a lagged three-week moving average. The figure highlights pronounced wave-like behavior and temporal non-stationarity in outcome prevalence, providing empirical justification for temporal validation, probability calibration, and threshold-based decision analysis in subsequent modeling stages.

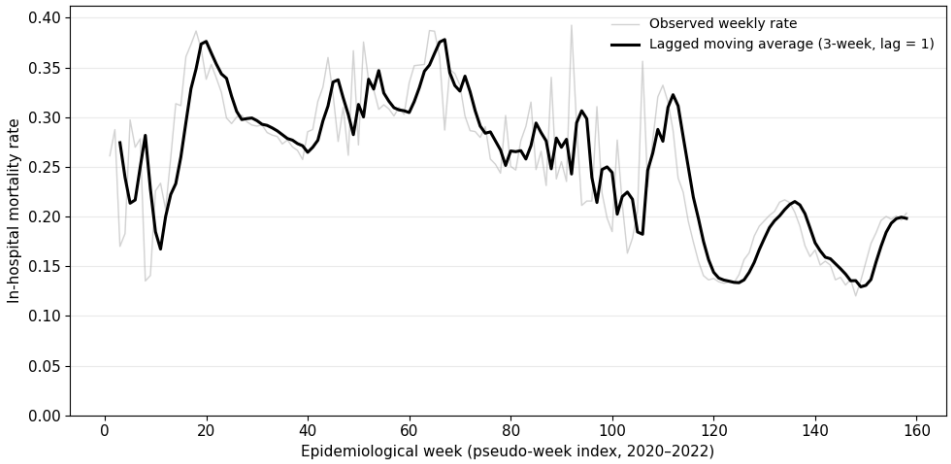

Figure 1 – Weekly in-hospital mortality rate illustrating temporal non-stationarity and leakage-safe smoothing using a lagged 3-week moving average (lag = 1)

Discrimination and calibration performance

Model performance was evaluated on the independent 2022 temporally held-out test set using metrics reflecting both discriminative ability and probability accuracy. Evaluation was conducted for individual base learners, the ensemble model, and their calibrated variants using a unified and fixed evaluation protocol.

Across all evaluated configurations, ensemble aggregation consistently yielded higher discriminative performance compared to single-model counterparts. Individual base learners demonstrated comparable but lower AUROC and PR-AUC values, whereas ensemble predictions achieved more stable performance across evaluation runs. This pattern was consistent across both raw and calibrated probability outputs.

For the final ensemble model, post-processing with isotonic calibration preserved discriminative performance while improving probability accuracy. On the 2022 test set, the isotonic-calibrated ensemble achieved an AUROC of 0.812, a PR-AUC of 0.483, and a Brier score of 0.140. Compared to the uncalibrated ensemble, calibration resulted in a reduction of the Brier score, indicating improved

alignment between predicted probabilities and observed outcomes, while AUROC and PR-AUC values remained largely unchanged.

When comparing raw and calibrated probability estimates, discrimination metrics (AUROC and PR-AUC) exhibited minimal variation across model variants. In contrast, probability accuracy differed more substantially, with calibrated outputs consistently demonstrating lower Brier scores than their uncalibrated counterparts. This pattern was observed both for individual base learners and for the ensemble model.

Comparative analysis between single models and the ensemble further indicated that ensemble predictions demonstrated more favorable combinations of discrimination and reliability metrics. While individual models achieved reasonable AUROC values, their probabilistic accuracy was generally inferior to that of the ensemble, particularly after calibration. These results suggest that ensemble aggregation and calibration contribute complementary improvements to overall model performance.

A quantitative summary of discrimination and reliability metrics for individual models, the ensemble, and their calibrated variants is presented in Table 1, enabling direct comparison of performance across evaluated configurations.

Table 1 – Discrimination and calibration performance of individual models, ensemble model, and calibrated ensemble on the temporally held-out 2022 test set

Model	AUROC	PR-AUC	Brier
LR	0.786	0.451	0.146
LR (Elastic Net)	0.770	0.428	0.150
RF	0.784	0.437	0.145
HGB	0.793	0.453	0.141
LGBM	0.793	0.453	0.143
XGB	0.792	0.453	0.143
Ensemble (raw)	0.802	0.475	0.141
Ensemble (isotonic)	0.812	0.483	0.140

Calibration assessment

Calibration performance was evaluated on the independent, temporally held-out 2022 test set by comparing predicted probabilities with observed outcome frequencies using calibration curves and probabilistic error measures. Calibration analysis was conducted for individual base learners, the ensemble model, and their post-processed variants obtained via isotonic regression.

Calibration curves indicated that the ensemble model exhibited good probability alignment already in its raw (uncalibrated) form, particularly within the clinically relevant range of predicted probabilities. While minor local deviations from the reference diagonal were observed, no evidence of systematic miscalibration was identified for ensemble predictions.

Post-hoc isotonic calibration preserved this overall alignment and resulted in modest but consistent improvements in probability reliability. For the ensemble model, isotonic calibration led to a reduction in the Brier score relative to the uncalibrated variant, while discrimination metrics remained largely unchanged. Similar calibration patterns were observed for individual base learners, although the magnitude of calibration effects varied across model families.

Comparative analysis further demonstrated that ensemble predictions exhibited smaller overall calibration deviations than single-model predictions, both before and after calibration. Post-calibration curves for the ensemble were smoother and remained closely aligned with observed outcome frequencies, whereas calibration curves for individual models showed greater residual variability.

Representative calibration curves for raw and isotonic-calibrated ensemble predictions are shown in Figure 2, illustrating probability alignment within the operational risk range and the stabilizing effect of post-hoc calibration.

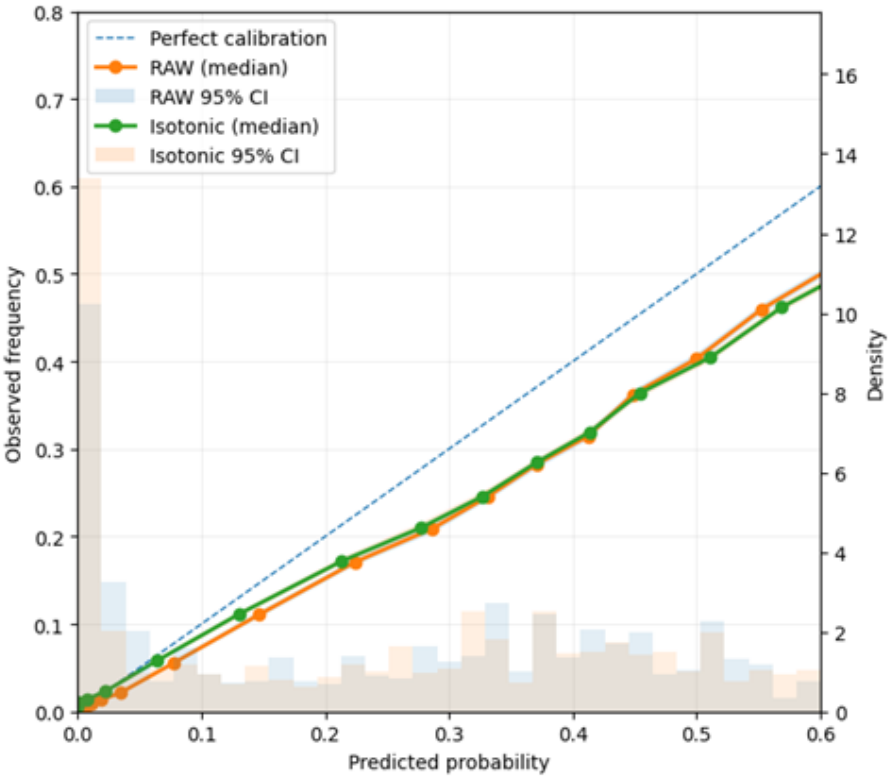

Figure 2 – Calibration of ensemble predictions on the 2022 test set using quantile-based binning with 95% bootstrap confidence intervals

Decision Curve Analysis

Decision Curve Analysis (DCA) was conducted on the independent 2022 temporally held-out test set to evaluate the clinical utility of the final isotonic-calibrated ensemble model across a range of probability thresholds. Net benefit curves were compared against the default reference strategies “treat all” and “treat none”.

Across the evaluated threshold range, the isotonic-calibrated ensemble demonstrated consistently positive net benefit relative to both reference strategies, particularly within the clinically plausible threshold interval of approximately 0.15 to 0.50. Within this range, the model-assisted decision strategy remained stably superior to baseline approaches, indicating a favorable balance between true-positive identification and avoidance of unnecessary interventions. This pattern indicates that the observed decision-analytic gains are primarily attributable to improved probability reliability rather than changes in ranking performance.

At lower thresholds, net benefit values approached those of the “treat all” strategy, reflecting increasingly liberal decision policies, whereas at higher thresholds the net benefit gradually converged toward the “treat none” strategy. This behavior is consistent with expected decision-analytic dynamics under increasing decision stringency.

Overall, these results indicate that the calibrated ensemble model provides meaningful clinical benefit across a broad range of early triage thresholds. Net benefit curves for the evaluated strategies are shown in Figure 3.

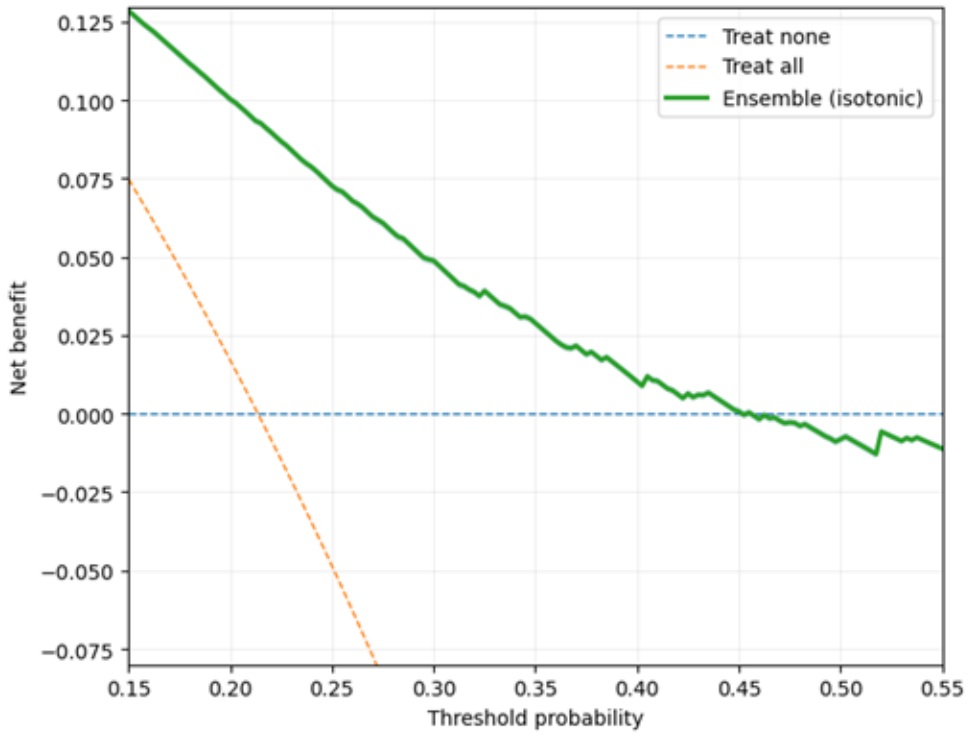

Figure 3 – Decision curve analysis for the isotonic-calibrated ensemble model on the independent 2022 test set, showing net benefit across threshold probabilities in comparison with default strategies (“treat all” and “treat none”)

Operational risk thresholds

To complement decision-curve–based evaluation, model performance was additionally examined at a set of operational probability thresholds relevant for early clinical triage. Thresholds were selected within the region where the calibrated ensemble demonstrated sustained positive net benefit in Decision Curve Analysis, focusing on values commonly considered in early risk stratification settings.

Three representative thresholds–0.20, 0.30, and 0.35–were evaluated on the independent 2022 temporally held-out test set. For each threshold, standard classification metrics were computed, including sensitivity, specificity, and positive predictive value, alongside corresponding net benefit estimates. This threshold-based evaluation enables direct comparison of model behavior across different operating points.

At lower thresholds, the calibrated ensemble identified a larger proportion of high-risk patients, resulting in higher sensitivity and lower specificity. As the threshold increased, sensitivity decreased while specificity increased, reflecting the expected trade-off between false-positive and false-negative decisions. Across all evaluated thresholds, the calibrated ensemble maintained positive net benefit relative to baseline strategies.

Comparative analysis indicated that threshold-based performance of the calibrated ensemble was more stable than that of uncalibrated predictions. In particular, calibrated outputs exhibited smaller fluctuations in net benefit across adjacent thresholds, whereas uncalibrated outputs showed greater sensitivity to threshold selection. This pattern was consistent with the calibration improvements observed for the ensemble model.

A summary of threshold-specific performance metrics for the calibrated ensemble is provided in Table 2, including sensitivity, specificity, positive predictive value, and net benefit at each evaluated

threshold. These results illustrate how calibrated probability estimates support consistent model behavior across a range of practical operating points.

Table 2 – Performance of the isotonic-calibrated ensemble model at selected probability thresholds on the 2022 test set

Threshold	Sensitivity	Specificity	PPV	Net benefit
0.20	0.922	0.508	0.337	0.1002
0.30	0.814	0.629	0.373	0.0488
0.35	0.723	0.703	0.398	0.0289

This study investigated early-stage machine learning models for in-hospital mortality prediction under conditions closely approximating real-world clinical deployment. By combining strict temporal validation, probability calibration, and decision-analytic evaluation, the results demonstrate that predictive discrimination alone is insufficient for reliable early clinical decision support, particularly in non-stationary settings such as infectious disease outbreaks [3–4].

Consistent with prior research, ensemble aggregation outperformed individual base learners in terms of discriminative performance. Ensemble methods are known to reduce variance and improve robustness when applied to large-scale observational healthcare data. However, the present findings show that improved discrimination does not automatically translate into clinically reliable risk estimates. Despite comparable AUROC and PR-AUC values, uncalibrated models exhibited local deviations in probability alignment, highlighting the limitations of relying solely on ranking-based metrics in threshold-dependent clinical applications [3].

Probability calibration emerged as a critical determinant of downstream model behavior. Calibration assessment and Brier score analysis demonstrated that isotonic post-processing resulted in modest but consistent improvements in agreement between predicted risks and observed outcomes without altering the underlying ranking structure. Together, these results reinforce the need to explicitly evaluate probability reliability when model outputs are used to guide early triage decisions, where absolute risk estimates directly influence threshold selection and resource allocation [9, 18].

The clinical relevance of probability calibration was further illustrated through Decision Curve Analysis. The isotonic-calibrated ensemble demonstrated consistently positive net benefit relative to default decision strategies across a broad range of clinically plausible thresholds. This finding indicates that improvements in probability reliability translate into more robust decision-analytic performance, rather than merely improved statistical metrics. By framing evaluation in terms of net benefit, Decision Curve Analysis provides a transparent link between predictive modeling and clinical decision-making and complements conventional discrimination-based assessment [10–11].

A key methodological strength of this study is the use of strict temporal validation. Training models on earlier pandemic waves and evaluating them on a fully independent, later time period revealed substantial shifts in outcome prevalence and baseline risk distributions. Under such conditions, commonly used random-split or cross-validation approaches may produce overly optimistic performance estimates. The persistence of calibration quality and decision-analytic benefit under temporal generalization suggests that the proposed framework offers a more realistic assessment of model robustness in evolving clinical environments [14].

The study also demonstrates that meaningful early risk stratification can be achieved using an EHR-lite feature set consisting exclusively of information available at hospital admission. While richer longitudinal, laboratory, or imaging data may further enhance predictive accuracy, such data are often unavailable at the time when early clinical decisions must be made. Similar observations have been reported in prior large-scale evaluations of clinical prediction models, underscoring the practical value of early-stage approaches under constrained data availability [4].

Several considerations are important when interpreting the results of this study. The analysis is based on a large-scale secondary clinical registry, which provides a realistic representation of routine hospital data but may limit direct generalizability to healthcare systems with different patient populations and data collection practices. At the same time, the use of strict temporal validation allows for a more realistic assessment of model performance under temporal distribution shift, which is critical for real-world deployment. In addition, while the present study focuses on fixed out-of-time evaluation, real-world clinical use may require ongoing monitoring of probability calibration as data distributions evolve over time. Future work may further explore adaptive recalibration and model updating strategies to maintain the stability and reliability of probabilistic predictions [14].

The analysis focused on a single binary outcome. At the same time, preliminary subgroup-level evaluation indicated broadly consistent model behavior across major demographic groups. However, more detailed subgroup analysis, fairness assessment, and prospective evaluation in real-world clinical workflows remain important directions for future work.

Conclusion

This study demonstrated that in early-stage in-hospital mortality prediction, probability reliability and decision-oriented evaluation are essential complements to discriminative performance. While ensemble learning improved overall predictive accuracy, post-hoc probability calibration contributed to stabilizing risk estimates and supporting more consistent threshold-based decision-making.

Using strict temporal validation and an early-stage (EHR-lite) feature set, the results showed that calibrated ensemble models maintained robust performance under dataset shift and achieved consistently positive net benefit across a clinically plausible range of decision thresholds. These findings highlight the importance of integrating calibration and decision-analytic methods into clinical machine learning pipelines, particularly in settings characterized by temporal non-stationarity and limited early information.

Overall, the proposed framework provides a practical and methodologically sound approach for evaluating early clinical risk prediction models and may support the development of more reliable machine learning–based decision-support tools in real-world hospital settings.

REFERENCES

1 Al-Nafjan, A., Aljuhani, A., Alshebel, A., Alharbi, A., Alshehri, A. Artificial intelligence in predictive healthcare: A systematic review. Journal of Clinical Medicine, 14 (19), 6752 (2025). https://doi.org/10.3390/jcm14196752

2 Bajwa, J., Munir, U., Nori, A., Williams, B. Artificial intelligence in healthcare: Transforming the practice of medicine. Future Healthcare Journal, 8 (2), e188–e194 (2021). https://doi.org/10.7861/fhj.2021-0095

3 Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195 (2019). https://doi.org/10.1186/s12916-019-1426-2

4 Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., COVID-19 Prediction Model Consortium. Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. BMJ, 369, m1328 (2020). https://doi.org/10.1136/bmj.m1328

5 Rajwa, B., Naved, M. M. A., Adibuzzaman, M., Grama, A. Y., Khan, B. A., Dundar, M., Rochet, J.-C. Identification of predictive patient characteristics for assessing the probability of COVID-19 in-hospital mortality. PLOS Digital Health, 3 (4), e0000327 (2024). https://doi.org/10.1371/journal.pdig.0000327

6 Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85 (2022). https://doi.org/10.1214/21-SS133

7 Ennab, M., Mcheick, H. Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions. Frontiers in Robotics and AI, 11, 1444763 (2024). https://doi.org/10.3389/frobt.2024.1444763

8 Tonekaboni, S., Joshi, S., McCradden, M. D., Goldenberg, A. What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of Machine Learning Research, 106, 359–380 (2019). https://proceedings.mlr.press/v106/tonekaboni19a.html

9 Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W. Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230 (2019). https://doi.org/10.1186/s12916-019-1466-7

10 Vickers, A. J., Van Calster, B., Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ, 352, i6 (2016). https://doi.org/10.1136/bmj.i6

11 Sadatsafavi, M., Adibi, A., Puhan, M., Gershon, A., Aaron, S. D., Sin, D. D. Moving beyond AUC: Decision curve analysis for quantifying net benefit of risk prediction models. European Respiratory Journal, 58 (5), 2101186 (2021). https://doi.org/10.1183/13993003.01186-2021

12 Andaur Navarro, C. L., Damen, J. A. A., van Smeden, M., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., Hooft, L. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. Journal of Clinical Epidemiology, 154, 1–12 (2023). https://doi.org/10.1016/j.jclinepi.2022.11.015

13 Giddings, R., Joseph, A., Callender, T., Janes, S. M., van der Schaar, M., Sheringham, J., Navani, N. Factors influencing clinician and patient interaction with machine learning-based risk prediction models: A systematic review. The Lancet Digital Health, 6 (2), e131–e144 (2024). https://doi.org/10.1016/S2589-7500(23)00241-8

14 Subbaswamy, A., Saria, S. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. Biostatistics, 21 (2), 345–352 (2020). https://doi.org/10.1093/biostatistics/kxz041

15 Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G. Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31 (12), 2346–2363 (2019). https://doi.org/10.1109/TKDE.2018.2876857

16 dos Santos Silva, G. F., Barcellos Filho, F. N., Wichmann, R. M., da Silva Junior, F. C., Chiavegatto Filho, A. D. P. Strategies for detecting and mitigating dataset shift in machine learning for health predictions: A systematic review. Journal of Biomedical Informatics, 170, 104902 (2025). https://doi.org/10.1016/j.jbi.2025.104902

17 Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., Collins, G. S. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. BMJ, 384, e074820 (2024). https://doi.org/10.1136/bmj-2023-074820

18 Austin, P. C., Steyerberg, E. W. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine, 38 (21), 4051–4065 (2019). https://doi.org/10.1002/sim.8281

19 Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., Ziegler, A. Calibrating machine learning approaches for probability estimation: A comprehensive comparison. Statistics in Medicine, 42 (29), 5451–5478 (2023). https://doi.org/10.1002/sim.9921

20 Gneiting, T., Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102 (477), 359–378 (2007). https://doi.org/10.1198/016214506000001437

21 Chalkou, K., Vickers, A. J., Pellegrini, F., Manca, A., Salanti, G. Decision curve analysis for personalized treatment choice between multiple options. Medical Decision Making, 43 (3), 337–349 (2022). https://doi.org/10.1177/0272989X221143058

22 Piovani, D., Sokou, R., Tsantes, A. G., Vitello, A. S., Bonovas, S. Optimizing clinical decision making with decision curve analysis: Insights for clinical investigators. Healthcare, 11 (16), 2244 (2023). https://doi.org/10.3390/healthcare11162244

[1]*Куничик В.Г.,
докторант, ORCID ID: 0009-0004-8580-7731,
*e-mail: kunichikval@gmail.com
[1]Салыкова О.С.,
т.ғ.к., қауымдастырылған профессор, ORCID ID: 0000-0002-8681-4552,
e-mail: solga0603@mail.ru

[1]Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті,
Қостанай қ., Қазақстан

ИНТЕРПРЕТАЦИЯЛАНАТЫН ЖӘНЕ КАЛИБРЛЕНГЕН АНСАМБЛЬДІК МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН ҚОЛДАНА ОТЫРЫП АУРУХАНА ІШІНДЕГІ ӨЛІМ-ЖІТІМДІ ЕРТЕ БОЛЖАУ

Аңдатпа

Аурухана жағдайында пациенттің өлім-жітім қаупін ерте кезеңде бағалау клиникалық тәжірибедегі маңызды міндеттердің бірі, әсіресе COVID-19 сияқты жұқпалы аурулар кең таралған жағдайда. Мұндай кезде шешімдер көбіне науқасты стационарға қабылдау сәтінде, шектеулі бастапқы клиникалық мәліметтер негізінде қабылданады. Машиналық оқыту әдістері бұл бағытта кеңінен қолданылғанымен, көптеген зерттеулер ерте клиникалық триаж талаптарына толық сәйкес келмейді. Бұл ең алдымен, ықтималдық болжамдарының сенімділігіне, уақыт бойынша өзгерістерге бейімделуіне және шешім қабылдауға негізделген бағалау тәсілдеріне жеткілікті көңіл бөлінбеуімен түсіндіріледі. Осы жұмыста ауруханаға түскен сәтте қолжетімді күнделікті клиникалық деректер негізінде өлім-жітім қаупін ерте болжау мәселесі қарастырылады. Зерттеу барысында нақты клиникалық жағдайларды бейнелеу мақсатында деректер уақыт бойынша қатаң бөлініп, модельдер осы қағидаға сәйкес оқытылып, бағаланды. Градиенттік бустингке негізделген ансамбльдік модельдер құрылып, болжамдардың сенімділігін арттыру үшін post-hoc изотоникалық калибрлеу әдісі қолданылды. Модельдердің тиімділігі дискриминациялық және калибрлеу метрикалары арқылы кешенді бағаланып, олардың клиникалық маңыздылығы шешім қабылдау қисықтарын талдау (Decision Curve Analysis, DCA) әдісімен зерттелді. Тәуелсіз, уақыт бойынша бөлінген тестілік жиында алынған нәтижелер ансамбльдік тәсілдің тәуекел бағаларының тұрақтылығын арттыратынын, ал калибрлеудің модельдің шекті мәндердегі жұмысын тұрақтандыратынын көрсетті. Сонымен қатар, калибрленген болжамдардың стандартты стратегиялармен салыстырғанда шекті мәндердің кең ауқымында жоғары тиімділік көрсететіні анықталды. Алынған нәтижелер ерте клиникалық шешім қабылдау жағдайында ықтималдық болжамдарының сенімділігі дискриминациялық қабілетпен қатар маңызды рөл атқаратынын дәлелдейді. Ұсынылған тәсіл уақыт бойынша өзгеретін деректер жағдайында және бастапқы ақпарат шектеулі болған кезде ауруханаішіндегі өлім-жітім қаупін бағалауға арналған сенімді әрі практикалық тұрғыдан негізделген әдіс.

Түйін сөздер: машиналық оқыту, ауруханаішіндегі өлім-жітім, ерте клиникалық триаж, ансамбльдік модельдер, ықтималдықтарды калибрлеу, шешім қабылдау қисықтарын талдау, клиникалық шешімдерді қолдау.

[1]*Куничик В.Г.,
докторант, ORCID ID: 0009-0004-8580-7731,
*e-mail: kunichikval@gmail.com
[1]Салыкова О.С.,
к.т.н., ассоциированный профессор,
ORCID ID: 0000-0002-8681-4552, e-mail: solga0603@mail.ru

[1]Костанайский региональный университет им. Ахмета Байтурсынулы,
г. Костанай, Казахстан

РАННЕЕ ПРОГНОЗИРОВАНИЕ ВНУТРИБОЛЬНИЧНОЙ ЛЕТАЛЬНОСТИ С ИСПОЛЬЗОВАНИЕМ ИНТЕРПРЕТИРУЕМЫХ И КАЛИБРОВАННЫХ АНСАМБЛЕВЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Аннотация

Раннее выявление пациентов с высоким риском внутрибольничной летальности остается одной из ключевых задач клинической практики, особенно в условиях вспышек инфекционных заболеваний, таких как COVID-19, когда решения должны приниматься уже на этапе госпитализации при ограниченном объеме доступной информации. Несмотря на широкое применение методов машинного обучения для прогнозирования летальности, многие существующие модели разрабатываются и оцениваются в условиях, снижающих их применимость для ранней клинической триаж-оценки, включая недостаточное внимание к надежности вероятностных прогнозов, временной обобщаемости и ориентированной на принятие решений оценке. В работе рассматривается задача раннего прогнозирования риска летального исхода на основе рутинно доступных клинических данных, собираемых при поступлении пациента в стационар. Разрабатывается ансамбль моделей градиентного бустинга, анализируемый в условиях строгого временного разделения обучающей и тестовой выборок, что отражает реалистичные сценарии практического внедрения. Особый акцент сделан на надежности прогнозируемых вероятностей: для улучшения согласованности между предсказанными рисками и наблюдаемыми исходами применяется пост-хок изотоническая калибровка. Качество моделей оценивается с использованием совокупности дискриминационных и калибровочных метрик, а клиническая значимость анализируется с помощью анализа кривых принятия решений в диапазоне практически значимых порогов. Оценка на независимой временно отложенной тестовой выборке показывает, что ансамблевая агрегация повышает устойчивость оценок риска, а калибровка обеспечивает более согласованное пороговое поведение без изменения ранжирующей способности моделей. Анализ кривых принятия решений демонстрирует преимущество откалиброванных прогнозов по чистому выигрышу в широком диапазоне порогов раннего клинического триажа. Полученные результаты подчеркивают, что надежность вероятностных оценок играет ключевую роль наряду с дискриминационной способностью моделей. Предложенный подход представляет собой методологически обоснованное решение для ранней оценки риска внутрибольничной летальности в условиях реальной клинической практики и временной нестабильности данных.

Ключевые слова: машинное обучение, внутрибольничная летальность, ранний клинический триаж, ансамблевые модели, калибровка вероятностей, анализ кривых принятия решений, поддержка клинических решений.