Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.37859

Dynamic Risk Prediction of Patient Clinical Deterioration Using Time-Aware Explainable Machine Learning

Yutong Li

Department of Biostatistics, School of Public Health, City University of Hong Kong, Hong Kong, China
liyutong02571@outlook.com

Abstract. The early identification of clinical instability is important in intensive care because such clinical instability usually occurs along a trajectory of abnormalities but not just one abnormal value. The authors designed a time-aware explainable machine learning system that calculates every four hours the deterioration risk based on the previous 24 hours' electronic health records and predicts a new invasive mechanical ventilation, administration of vasopressors, or death in 12 hours. Time-varying vital signs, lab results, treatments, frequency, and missingness are converted into level, variability, trend, change, and recency features and analyzed using LightGBM. The treeSHAP algorithm connects subsequent predictions with changes in the patient's risk factors. The internal performance was AUROC 0.901 ± 0.005 and AUPRC 0.428 ± 0.013 in contrast to latest-value LightGBM AUROC 0.867 ± 0.006 and AUPRC 0.356 ± 0.012. In external validation, the authors obtained AUROC 0.872 ± 0.006, Brier Score 0.078 ± 0.003, and warning lead time 7.6 ± 2.9 h. Dynamic SHAP analysis revealed a pattern that the increased respiratory rate, decreasing mean arterial pressure, elevated lactate, and neurological impairment are common before a prolonged increase in risk.

Keywords: clinical deterioration, dynamic risk prediction, electronic health records, LightGBM, explainable machine learning

1. Introduction

Clinical deterioration on the ICU can develop via respiratory failure, circulatory instability, neurologic deterioration, or multiorgan dysfunction before an adverse event becomes manifest. Recognition thus becomes a longitudinal inferencing challenge since clinicians must determine the distinction between transient physiological anomalies and persisting and coherent changes in several physiological systems. It has been increasingly recognized that real-time machine-learning-based early warning systems can affect escalation routes since predictions have become part of clinical workflows [1]. Prospective implementation of the Targeted Real-time Early Warning System for sepsis has shown that continuously updating electronic health record features can be utilized for earlier detection and treatment rather than just risk stratification [2]. Real-time machine-learning alerts have also been prospectively tested in a clinical setting, underscoring the necessity of investigating how early prediction leads are for affecting management [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.37859

Traditional warning scores are still valuable due to their simplicity and transparency; however, fixed thresholds convert the series of measurements into a current-state feature. Consequently, they do not reflect the rate of physiological change or the irregularities of observations' times. Machine learning can learn complex interactions of the current state, recently observed variability, context of treatments and temporality of changes; nevertheless, predictive accuracy alone is not enough. Reviews of clinical machine-learning trials show that even methodologically sound performance does not guarantee clinically useful interventions [4]. Another important issue is generalizability: externally validated intensive-care prediction models generally suffer from a loss of discrimination when applied in new institutions due to the differences in patient mix, measurement practices, and documentation system [5].

This research proposes a time-aware and explainable framework for dynamic prediction of clinical deterioration. Risk is reassessed every four hours based on the leakage-controlled 24-hour observation window and a 12-hour deterioration horizon. LightGBM learns the interaction of temporal summaries and measurement recency, whereas TreeSHAP decomposes the prediction and the difference with the previous one. MIMIC-IV is used for model development and internal validation, whereas eICU is used for external validation. The framework evaluates the discriminatory ability, calibration, warning lead time, and physiological contributors to the change in the patient-level risk.

2. Literature review

2.1. Dynamic prediction of clinical deterioration

Time series electronic health records make deterioration modeling shift from admission-level prognostic task to dynamic landmarking with repeated predictions. Time-structured MIMIC-IV ICU data provide a possibility to reconstruct the historical windows without using future information [6]. Quality of such models highly depends on data quality due to different levels of missingness, plausibility of measurements, and sampling rates among other variables [7]. Thus, dynamic prediction needs preservation of both physiological history and observation timing. Repeated estimation of risk alters the evaluation of such models where stable discrimination, calibrated probabilities, and timely alerts need to be ensured. Therefore, the need for leakage control and individual-level validation is higher than in usual admission-level prediction.

2.2. Explainability in clinical machine learning

Explainability is especially important in case of dynamic alerting since it concerns two things at once – why the patient is high risk and why this risk changed. SHAP-based ICU models proved that the nonlinear ensemble predictions can be decomposed into contributions made by population-level features and by individual patient features [8]. Recent researches combining multi-timepoint risk prediction, external validation, and SHAP interpretation of LightGBM models showed the importance of linking model performance to clinically understandable risk factors [9]. Likewise, continuous physiological prediction studies focus on the dynamic aspect of assessing deterioration [10]. Thus, consecutive SHAP vectors connection will allow differentiating between persistent baseline vulnerability and newly appearing factors and whether an increased risk was driven by deterioration of respiratory, circulatory, metabolic, renal, or neurological parameters.

DOI: 10.54254/2755-2721/2027.37859

3. Methodology

3.1. Data sources and cohort construction

The development cohort was drawn from MIMIC-IV, while the external cohort was extracted from the eICU Collaborative Research Database. Patients aged ≥18 years with a stay in the ICU that lasted more than 24 hours and whose time-stamped vital signs data could be accessed were included. Only the first qualifying ICU admission during a hospitalization was used. ICU admissions without outcome data that could be used were excluded. PostgreSQL enabled extraction, and Python 3.11 was used in data processing. Variables in the shared variable space included demographics, comorbidity burden, heart rate, respiratory rate, blood pressure, oxygen saturation, temperature, Glasgow Coma Score, complete blood count, electrolyte levels, kidney function biomarkers, blood gases, lactate levels, and treatment information. The overall time-aware prediction and explanation framework is illustrated in Figure 1, integrating longitudinal ICU data, dynamic temporal feature construction, LightGBM-based risk estimation, and TreeSHAP-based temporal interpretation.

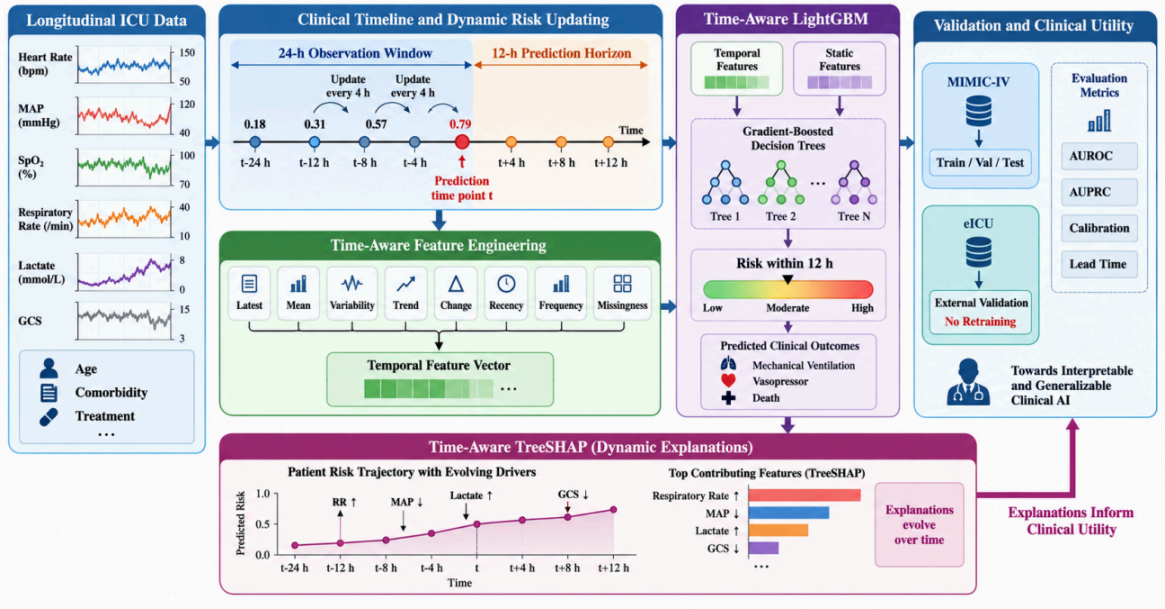


Figure 1. Time-aware explainable machine learning framework for dynamic clinical deterioration prediction

3.2. Dynamic prediction window and outcome

For each ICU admission, the periods were landmarked every four hours starting at 24 hours after admission to the ICU. For any given landmark t, information was available from the previous 24 hours, and the aim was to predict deterioration in the next 12 hours. The follow-up periods were terminated on either discharge, occurrence of deterioration, or end of observation time. Deterioration was taken as the occurrence of invasive mechanical ventilation, vasopressor administration, and in-hospital mortality. Invasive mechanical ventilation and vasopressor administrations were considered fresh only if there was no such support in the previous period. Any window after the first deterioration was disregarded. Since multiple windows of one patient were available, the patient-based division was done prior to developing the models so that the observations of one patient do not fall into two different sets.

DOI: 10.54254/2755-2721/2027.37859

3.3. Time-aware feature engineering

In regard to each repeated measure variable, the 24 hour period yielded the most recent value, average, lowest, highest, standard deviation, change between the first and last values, number of measurements, time since the most recent measurement, indicator of missing data, and a slope weighted by recency. Such a design differentiates between present physiological state and trend, instability, and rate of observations. The more recent the observations were, the higher was their impact on calculating the trend. The dynamic risk for patient i at time t was defined as shown in Equation (1):

$$\widehat{p}_i(t) = \sigma[\sum_{m=1}^{M} \eta f_m(s_i, \{\ell_{ij}, \mu_{ij}, \sigma_{ij}, \Delta_{ij}, \log(1+n_{ij}), e^{-\tau_{ij}/\tau_j}, q_{ij}, \beta_{ij}\}_{j=1}^{J}) \quad (1)$$

\Here, s contains static characteristics, l is the latest value, μ and σ describe level and variability, Δ is net change, n is measurement frequency, r is recency, q denotes missingness, and β represents recency-weighted trend.

3.4. LightGBM risk prediction model

LightGBM learned the heterogeneous matrix of features. Optuna chose the learning rate, number of leaves, tree depth, minimal number of samples in a leaf, feature fraction, bagging fraction, and regularization via maximization of the mean AUPRC on five folds with stratified patients' grouping. The class imbalance was addressed via inverse frequency weighting without artificial oversampling, while early stopping was used to manage the length of boosting. Once hyperparameters and threshold were fixed based on validation, the model returned calibrated probabilities every four hours.

3.5. Time-aware explainability

Feature attributions were done for every prediction made by the LightGBM model using TreeSHAP. Mean absolute SHAP was calculated for global importance. Features derived from the same physiological feature were combined together for easier interpretation by clinicians. Patient SHAP vectors were sorted chronologically, and differences between adjacent vectors allowed identification of features contributing to risk change. A new driver needed to have an increasing contribution to risk according to the total risk predicted in adjacent landmarks.

4. Experimental design and validation

4.1. Comparative experimental design

MIMIC-IV patients were stratified at the patient-level into 70%, 15%, and 15% for training, validation, and internal test sets, respectively. The proposed time-sensitive LightGBM algorithm was benchmarked against NEWS2, logistic regression, and recent LightGBM on only the latest measurements. Stratifications were identical for all models. Ablation tests eliminated trend, variability, and recency-based features to quantify their added value above nonlinear boosting. Preprocessing parameters, class weights, hyperparameters, and threshold were chosen using only the training and validation datasets without knowledge of the internal test set.

DOI: 10.54254/2755-2721/2027.37859

4.2. Performance and external validation

AUROC and AUPRC quantified discrimination; sensitivity, specificity, F1 score, and positive predictive value were computed based on the selected threshold in validation data. Brier score was applied for calibration, and warning effectiveness was measured using the duration from the first high-risk alert until the event. Bootstrapping with 1,000 repetitions was performed using patient-level sampling to account for within-patient dependencies. The final MIMIC-IV model was deployed in eICU without re-training, tuning of the threshold, or recalibration.

5. Results

5.1. Dynamic prediction performance and external validation

MIMIC-IV contributed 42,816 eligible ICU stays and 518,604 prediction windows; eICU contributed 69,742 stays and 811,936 external windows. Time-aware LightGBM achieved internal AUROC 0.901 ± 0.005 and AUPRC 0.428 ± 0.013, versus 0.867 ± 0.006 and 0.356 ± 0.012 for latest-value LightGBM* (see Table 1). Its Brier score was 0.071 ± 0.003* and warning lead time 8.4 ± 2.7 h. External AUROC remained 0.872 ± 0.006 with AUPRC 0.381 ± 0.011. Removing trend, variability, and recency/frequency features reduced AUROC to 0.887 ± 0.006*, 0.892 ± 0.005*, and 0.884 ± 0.006*, respectively.

Table 1. Dynamic prediction and external validation results

Model / Setting	AUROC	AUPRC	Sensitivity	Specificity	F1-score	Brier Score	Lead Time (h)
NEWS2	0.748 ± 0.009	0.218 ± 0.012	0.711 ± 0.023	0.684 ± 0.021	0.311 ± 0.017	0.109 ± 0.005	4.7 ± 2.3
Logistic regression	0.821 ± 0.007	0.294 ± 0.011	0.758 ± 0.020	0.742 ± 0.018	0.362 ± 0.016	0.094 ± 0.004	5.6 ± 2.5
Latest-value LightGBM	0.867 ± 0.006	0.356 ± 0.012	0.801 ± 0.018	0.781 ± 0.017	0.412 ± 0.015	0.083 ± 0.004	6.5 ± 2.6
Time-aware LightGBM	0.901 ± 0.005*	0.428 ± 0.013*	0.843 ± 0.017*	0.812 ± 0.015*	0.468 ± 0.014*	0.071 ± 0.003*	8.4 ± 2.7*
External eICU	0.872 ± 0.006	0.381 ± 0.011	0.812 ± 0.018	0.789 ± 0.016	0.423 ± 0.015	0.078 ± 0.003	7.6 ± 2.9

**P < 0.05 after patient-level bootstrap comparison.*

5.2. Temporal risk explanations and clinical warning patterns

Among TreeSHAP contributions were recent lactate of 0.184 ± 0.073, minimum mean arterial pressure of 0.162 ± 0.068, respiratory rate slope of 0.145 ± 0.061, Glasgow coma score difference of 0.137 ± 0.057, and oxygen saturation fluctuation of 0.124 ± 0.052. The mean probability of deterioration increased from 0.231 ± 0.094 twelve hours prior to an event, to 0.418 ± 0.137 eight hours prior, 0.637 ± 0.154 four hours prior, and 0.781 ± 0.121 at the last landmark*. The warning lead time was 9.1 ± 2.4 hours prior to mechanical ventilation, 7.8 ± 2.6 h before vasopressor initiation, and 10.3 ± 3.1 h before death. Falling mean arterial pressure and rising lactate increased

DOI: 10.54254/2755-2721/2027.37859

log-odds contributions by 0.31 ± 0.12* and 0.28 ± 0.13*, supporting physiologically coherent risk escalation.

6. Conclusion

Time-aware models generate deterioration probabilities by translating longitudinal ICU data into constantly updated probability estimates, without violating direction, variability, recency, and sampling frequency of physiology. LightGBM achieved better discrimination and early warning time compared to baseline methods and performed consistently enough to allow external validation on eICU. Temporal ablation highlighted that temporal information had an added value beyond that of trends, variability, and recency. Consecutive TreeSHAP analysis allowed identifying coherent physiological changes leading to increased risk for each patient independently of a global measure of feature importance. Discrimination, calibration, warning time, external validation, and time-dependent explanations provide a clinically relevant approach to deterioration monitoring. Prospectively evaluating clinician behavior is needed.

References

[1] Rossetti, S. C., Dykes, P. C., Knaplund, C., et al. (2025). Real-time surveillance system for patient deterioration: A pragmatic cluster-randomized controlled trial. *Nature Medicine, 31*, 1895–1902.

[2] Adams, R., Henry, K. E., Sridharan, A., et al. (2022). Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine, 28*, 1455–1460.

[3] Levin, M. A., Kia, A., Timsina, P., Cheng, F. Y., Nguyen, K. A. N., Kohli-Seth, R., Lin, H. M., Ouyang, Y., Freeman, R., & Reich, D. L. (2024). Real-time machine learning alerts to prevent escalation of care: A nonrandomized clustered pragmatic clinical trial. *Critical Care Medicine, 52*(7), 1007–1020.

[4] Plana, D., Shung, D. L., Grimshaw, A. A., Saraf, A., Sung, J. J. Y., & Kann, B. H. (2022). Randomized clinical trials of machine learning interventions in health care: A systematic review. *JAMA Network Open, 5*(9), e2233946.

[5] Rockenschaub, P., Akay, E. M., Carlisle, B. G., et al. (2025). External validation of AI-based scoring systems in the ICU: A systematic review and meta-analysis. *BMC Medical Informatics and Decision Making, 25*, 5.

[6] Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data, 10*, 1.

[7] Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R. O., & Gupta, A. (2023). Electronic health record data quality assessment and tools: A systematic review. *Journal of the American Medical Informatics Association, 30*(10), 1730–1740.

[8] Hu, C., Gao, C., Li, T., Liu, C., & Peng, Z. (2024). Explainable artificial intelligence model for mortality risk prediction in the intensive care unit: A derivation and validation study. *Postgraduate Medical Journal, 100*(1182), 219–227.

[9] Ge, X., Chen, W., Shi, J., Zhang, J., Tai, H., Zhang, Y., Wang, B., Liu, W., Chen, S., & Han, H. (2025). Prediction of moderate-to-severe sepsis-associated acute kidney injury using a dual-timepoint machine learning model: Development, multiregional validation, and clinical deployment study. *Journal of Medical Internet Research, 27*, e73840.

[10] Park, C., Han, C., Jang, S. K., Kim, H., Kim, S., Kang, B. H., Jung, K., & Yoon, D. (2025). Development and validation of a machine learning model for early prediction of delirium in intensive care units using continuous physiological data: Retrospective study. *Journal of Medical Internet Research, 27*, e59520.