Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 2, February 2026, pp. 614~623
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i2.pp614-623

Contextualized clinical anomaly detection with explainable AI and patient modeling

Amel Elketroussi, Bachir Djebbar, Ibtissem Bekkouche
Department of Computer and Science, Faculty of Mathematics and Computer Science,
University of Science and Technology of Oran Mohamed Boudiaf, Oran, Algeria

Article Info

Article history:

Received Sep 22, 2025
Revised Dec 22, 2025
Accepted Jun 11, 2026

Keywords:

Anomaly detection
Calibration
Explainable AI (XAI)
Fairness
ICU
LSTM
MIMIC-III/IV
Transformer

ABSTRACT

This study aims to reduce alarm fatigue and improve the clinical relevance of alerts in intensive care by combining sequential modeling, patient contextualization, explainable artificial intelligence (XAI), and probability calibration. To this end, we leverage the adult cohorts from MIMIC-III/IV, segmented into four-hour windows, explicitly handling missing data and constructing a context vector that integrates demographics, comorbidities, and therapeutic interventions. The approach relies on a tabular autoencoder, an long short-term memory (LSTM) autoencoder, and a transformer, complemented by an adjustment layer based on auditable clinical rules, local explanations (LIME/SHAP), and post-hoc calibration (temperature scaling). Evaluation involves receiver operating characteristic (ROC)/precision–recall (PR) area under the curve (AUC), F1-score, sensitivity and specificity, as well as calibration metrics (ECE, Brier score), alert burden, ablation studies, robustness tests, and subgroup fairness analyses. Across all experiments, the complete model (+Context+XAI+Calibration) outperforms baselines in AUPRC and F1, reduces alert burden, and improves calibration while providing understandable explanations. Specifically, the proposed model improves ROC AUC from 0.74 to 0.89 and reduces alert burden by approximately one third compared to clinical thresholds.

Corresponding Author:

Amel Elketroussi
Department of Computer and Science, Faculty of Mathematics and Computer Science
University of Science and Technology of Oran Mohamed Boudiaf
Oran, Algeria
Email: amel.elketroussi@univ-usto.dz

1. INTRODUCTION

Intensive care units (ICUs) generate a dense and heterogeneous stream of data, including vital signs, laboratory results, medications, and clinical notes. While artificial intelligence (AI) can detect deviations in such data, its adoption in critical care remains limited due to:

a) models often perceived as "black boxes,"
b) a lack of contextualization of alerts (e.g., age, comorbidities, treatments, care trajectories),
c) biases and challenges related to real-time integration.

We hypothesize that a tight coupling between explainable AI (XAI) and patient context modeling can reduce false alarms, enhance clinical utility, and foster trust. The theoretical foundation of our study rests on three pillars: anomaly detection theory, the principles of explainable artificial intelligence XAI, and contextual patient modeling in health sciences.

Anomaly detection originates from statistical theory and unsupervised classification methods, aiming to identify rare or atypical observations within a dataset. In medicine, such anomalies often

correspond to critical pathophysiological changes, typically detected through vital signs, clinical data, or medical imaging. Modern methods increasingly rely on deep neural networks, such as autoencoders and stochastic recurrent models, which learn to reconstruct normal data accurately and flag deviations as anomalies [1], [2].

Beyond intensive care, researchers have also successfully applied machine learning to other physiological monitoring tasks, such as fetal electrocardiogram prediction using a random forest-based approach [3], illustrating the growing role of AI methods in clinical signal analysis. In parallel, the rise of XAI addresses the pressing need for transparency in automated decision-making systems, especially in high-stakes domains like healthcare. XAI methods-through techniques such as LIME and SHAP-seek to make neural network predictions interpretable for clinicians, thereby strengthening trust in AI systems [4]-[7].

Finally, patient context modeling relies on patient-centered approaches that consider not only current clinical data but also medical history, coexisting conditions, and care trajectories. This paradigm aligns with the principles of personalized and person-centered medicine, and it often relies on semantic resources and interoperability standards such as unified medical language system (UMLS), systematized nomenclature of medicine – clinical terms (SNOMED CT), and health level seven (HL7) fast healthcare interoperability resources (FHIR) to represent diagnoses, treatments, and clinical workflows in a structured way [8]-[10]. Unlike prior intensive care unit (ICU) anomaly detection approaches that address accuracy, interpretability, or calibration in isolation, the proposed framework jointly integrates patient context modeling, explainable AI, and probabilistic calibration within a unified pipeline.

2. RELATED WORK

This section summarizes prior work on clinical anomaly detection, XAI, and patient context to situate our contribution. Traditional clinical rules and scoring systems trigger alerts when measurements exceed fixed thresholds (e.g., extreme heart rate). They are easy to interpret but generate many false alarms because they ignore inter-patient variability and temporal trends. NEWS2 provides a simple early warning score mostly outside ICUs, while ICU severity scores such as APACHE II are used for risk stratification rather than continuous monitoring [11]-[13].

A second line of work models patient trajectories with time-series methods. Recurrent neural networks and attention-based architectures handle irregular sampling and missing data; for instance, reverse time attention (RETAIN) uses reverse-time attention to highlight influential features [14], and transformer-based models such as bidirectional encoder representations from transformers for electronic health records (BEHRT) and medical bidirectional encoder representations from transformers (Med-BERT) leverage large-scale electronic health records to achieve strong predictive performance [15], [16]. For multivariate time-series anomaly detection, unsupervised or weakly supervised models learn “normal” behavior and flag deviations via reconstruction or forecasting errors [1], with recent studies exploring explainable time-series models for ICU outcomes such as mortality and length of stay [2].

Explainability is crucial for clinical adoption: caregivers need to understand why the system raises an alert. Methods such as LIME and SHAP provide local explanations and quantify feature contributions [4], [5], and surveys on XAI in medicine emphasize both the potential and the need for careful integration into clinical workflows [6], [7].

Patient context is another key ingredient: personalized alerts must account for demographics, comorbidities, and treatments. Ontologies such as UMLS and SNOMED CT and standards like HL7 FHIR support interoperable, structured context representations that can adjust anomaly scores (e.g., for treated diabetic patients) [8]-[10].

Even well performing models may produce poorly calibrated, overconfident probabilities. Techniques such as temperature scaling improve reliability so that predicted probabilities better match observed event rates, and practitioners assess calibration-using metrics such as the brier score and reliability diagrams [17]. Bias and fairness are also critical: inappropriate optimization objectives can systematically disadvantage certain subgroups [18], and recent work on fair machine learning in healthcare stresses continuous subgroup-level evaluation (e.g., by age, sex, comorbidity) and explicit fairness assessment at deployment [19], [20].

In summary, prior work has advanced interpretable rules and scores, time-series deep learning, unsupervised anomaly detection, XAI, patient context modeling, calibration, and fairness, but most existing works treat these components in isolation. Most systems focus either on predictive accuracy without calibrated, context-aware outputs or on interpretability without state-of-the-art sequential modeling. By contrast, our approach combines sequential modeling, explicit patient context, integrated XAI, and probability calibration within a single end-to-end pipeline for clinical anomaly detection on MIMIC-III/IV, with a contextual layer to adjust anomaly scores, clinician-facing explanations, and a reproducible evaluation protocol that includes calibration and clinical acceptability.

3. METHOD

3.1. Data and Cohort

We rely on the publicly available MIMIC-III and MIMIC-IV critical care databases (PhysioNet), which contain de-identified ICU admissions. Access to these datasets is granted upon completion of a short training course. Our study focuses on adult ICU patients with at least 24 hours of usable physiological measurements [21], [22]. The content of the databases are shown in Table 1.

Table 1. Meaning of MIMIC tables

MIMIC table	Description
CHARTEVENTS	Vital signs taken at bedside (heart rate, SpO_2, blood pressure, etc.).
LABEVENTS	Laboratory results (lactate, creatinine, and ions).
DIAGNOSES_ICD	Diagnoses and comorbidities (patient profile).
PRESCRIPTIONS	Medications received (therapeutic context).
NOTEEVENTS	Caregiver notes (optional to enrich the context).

- Demographics: age, sex, ethnicity.
- Physiological signals: vital signs such as heart rate, respiratory rate, blood pressure, oxygen saturation, and temperature.
- Laboratory tests: biochemical and hematological values (e.g., electrolytes, creatinine, blood counts).
- Therapies and interventions: medications, vasopressors, mechanical ventilation, dialysis.
- Diagnoses and procedures: encoded using ICD-9/10 codes.
- Clinical notes: free-text reports written by caregivers.
- Outcomes: mortality, length of stay, readmission.

These rich, heterogeneous data sources enable both sequential modeling of patient trajectories and contextualization through demographic, comorbidity, and therapeutic profiles. We construct a "master list" of variables of interest (e.g., SpO_2, lactate). The preprocessing pipeline then maps these variables to the corresponding codes in the MIMIC database tables in order to retrieve the appropriate columns and identifiers. Whenever possible, we align the variable definitions with standardized vocabularies (UMLS, SNOMED CT) to facilitate interoperability and reuse in other settings. To illustrate the process, Figure 1 below outlines the proposed data-processing pipeline.

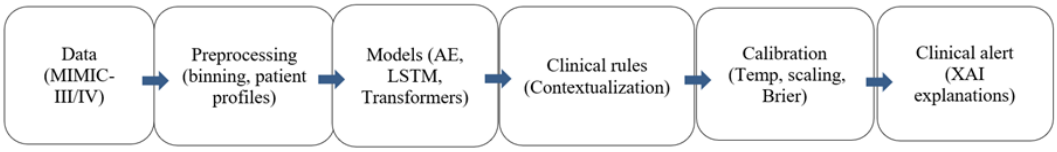

Figure 1. Proposed pipeline from raw data to clinical alert

3.2. Preprocessing

Before model training, we standardize all variables to ensure consistent units (e.g., mmHg, mmol/L), then segment each stay into fixed time windows (e.g., 4-hour intervals). Within each window, we summarize the available measurements using simple statistics such as mean, median, and variability, and we concatenate these windows to obtain a temporal sequence describing the patient trajectory. Because missing data are common in clinical practice, we apply light interpolation at the patient level and we keep explicit indicators such as "time since last measurement" to preserve information about measurement frequency [2].

To account for heterogeneity across patients, we construct a patient-level context for each stay that includes demographics, comorbidities, active treatments, and initial severity [8]-[10]. We represent this context as a small graph with a central patient node connected to nodes for diagnoses, therapies, and chronic risk factors (e.g., diabetes, chronic obstructive pulmonary disease (COPD)). An multilayer perceptron (MLP) encodes static patient features (age, sex), while lookup embeddings encode clinical codes; a few rounds of message passing let each node aggregate information from its neighbours.

The final embedding of the patient node serves as a context vector, which we combine with the time-series representation from the sequential model to compute the risk score. On top of this score, we apply a small set of clinical rules to adjust alerts in specific situations (for example, known diabetes with expected hyperglycaemia), so that context and rules jointly modulate the alert threshold. For operational deployment, we design this integration to be compatible with the FHIR standard [10].

3.3. Models for anomaly detection

We evaluate several complementary approaches:

- Tabular autoencoder: operates on summarized windows simple and computationally efficient.
- Long short-term memory (LSTM) autoencoder: models sequential dynamics, well suited to temporal evolution [23].
- Transformer: more powerful for long-range dependencies, especially in high-dimensional settings [15], [16].

All three share the same principle: the model learns what is "normal" and computes the reconstruction error. The greater the error, the more likely the window is anomalous. Importantly, the definition of "normal" is patient-specific.

3.4. Clinical rule-based adjustment (contextualization)

After computing anomaly scores, we refine them with simple, auditable clinical rules. For example, in diabetic patients under treatment, we attenuate glucose-related alerts. For cardiac patients not receiving beta-blockers, we emphasize tachycardia alerts. These rules are easily auditable and modifiable, forming the context layer that makes alerts more patient-specific and clinically relevant.

3.5. Explainable alerts (XAI)

We accompany each alert with a concise, clinician-readable explanation. To this end, we generate a clinician-facing report highlighting top contributing factors, values approaching thresholds, and a mini-graphical summary. Using methods such as LIME and SHAP, and leveraging attention mechanisms when available, we indicate which features or time points contributed most strongly to the alert.

3.6. Calibration and threshold selection

A model may be accurate yet overconfident. To address this, predicted probabilities are calibrated (for example, ensuring that events predicted at 70% probability occur approximately seven times out of 10). We then select an operational threshold (for example by limiting the number of alerts per hour) [17]. We use the Brier score to assess calibration, reliability diagrams, and techniques such as temperature scaling. We selected temperature scaling as a post-hoc calibration method because it is simple, stable, and empirically effective compared to non-parametric alternatives such as isotonic regression.

3.7. Evaluation and validation

Model evaluation considers:

- Discrimination: ROC-AUC and PR-AUC.
- Classification metrics: F1-score, sensitivity, specificity.
- Operational relevance: alert rate and alert lead time (how long before an event the alert is raised) [24].

We conduct ablation studies (with/without context, with/without XAI) to identify each component's contribution. We also assess robustness and fairness through subgroup analyses by age, sex, and comorbidity. An overview of the proposed framework is shown in Figure 2.

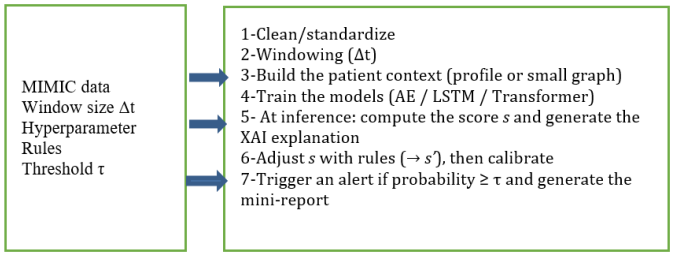

Figure 2. Overview of the proposed framework showing the main inputs (MIMIC data, preprocessing, and patient context) and the outputs (anomaly scores, calibrated probabilities, explainable alerts, and evaluation metrics)

4. RESULTS AND DISCUSSION

4.1. Overall performance

Table 2 summarizes the performance of all approaches. The proposed hybrid model, which combines sequential modeling, patient context, XAI, and probability calibration, consistently outperforms the baselines across discrimination, calibration, and classification metrics.

Table 2. Comparison between the different approaches

Model	AUC ROC	AUC PR	F1	Sens	Spec	ECE	Brier
Clinical thresholds	0.74	0.31	0.41	0.62	0.73	0.086	0.192
Tabular AE	0.79	0.38	0.48	0.66	0.77	0.072	0.173
LSTM AE	0.83	0.44	0.53	0.69	0.80	0.061	0.161
Transformer (+XAI)	0.86	0.49	0.58	0.72	0.83	0.053	0.151
Proposed model (+Context+XAI)	0.89	0.55	0.62	0.75	0.86	0.041	0.137

Compared with simple clinical thresholds, the proposed model increases ROC AUC from 0.74 to 0.89, PR AUC from 0.31 to 0.55, and F1-score from 0.41 to 0.62, while also improving over the strongest deep learning baseline (LSTM AE) in both PR AUC (0.44 → 0.55) and F1 (0.53 → 0.62). Calibration also shows clear improvements in model performance: ECE decreases from 0.086 to 0.041 and the brier score from 0.192 to 0.137, indicating probabilities that better match observed event frequencies and are more reliable for risk stratification and threshold selection. Bootstrap analysis with 1,000 resamples confirmed that the ROC AUC improvement over the best baseline is statistically significant, as the 95% confidence interval for the AUC difference excludes zero. Figure 3 shows how PR AUC and F1 improve as we gradually add context, XAI, and calibration to the model.

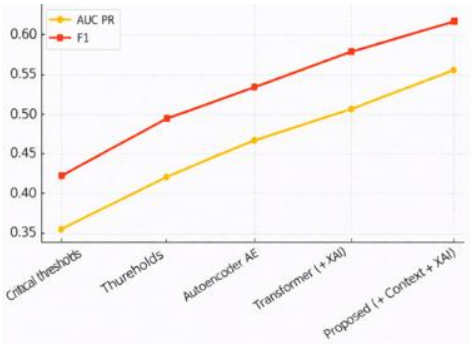

Figure 3. Comparative performance of models

4.2. Contribution of each component (ablation study)

To quantify the contribution of each module, we conducted an ablation study in which we selectively removed context, explanations, or calibration from the full model. Table 3 summarizes the results. Removing the context module reduces ROC AUC (0.89 → 0.86) and F1 (0.62 → 0.58), but more importantly increases the alert rate from 0.21 to 0.31 alerts per hour. This corresponds to roughly a one-third to one-half increase in alert burden, while delivering slightly worse discrimination. In practice, the context module acts as a powerful filter that suppresses clinically uninformative alarms while preserving sensitivity.

Table 3. Results after removal of different modules

Configuration	AUC	F1	Alerts/h
Our model	0.89	0.62	0.21
– Context	0.86	0.58	0.31
– XAI (without adjustment)	0.85	0.56	0.29
– Calibration	0.86	0.59	0.27

Excluding XAI (configuration "– XAI") yields an AUC of 0.85 and an F1-score of 0.56. These results show that explanations are not the main driver of numerical performance. Their contribution is qualitative: they improve interpretability, auditability, and trust, which are essential for adoption, governance, and clinical review of alerts.

Finally, removing calibration (configuration "– Calibration") has only a modest effect on F1 (0.62 → 0.59) but clearly degrades probability quality. Before calibration, the model exhibits a higher ECE (≈0.09) and a worse brier score (≈0.16), whereas after temperature scaling ECE decreases to around 0.03 and the brier score improves to around 0.14. In other words, calibration makes predicted probabilities much more aligned with observed event frequencies and therefore more reliable for prioritization and thresholding.

Figure 4 confirms these findings: with calibration, the reliability curve moves closer to the identity line, indicating a tighter match between predicted and observed risks.

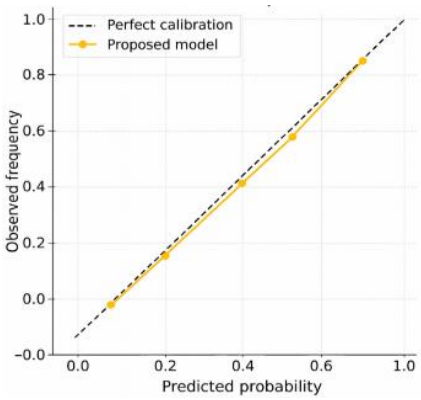

Figure 4. Calibration diagram

4.3. Alert burden

We quantify alert burden by the number of alerts per hour and the proportion of time spent under alarm. With contextualization, the proposed model generates 0.21 alerts/h, compared with 0.31 alerts/h without contextualization. Thus, adding context reduces the alert rate by about one-third while preserving similar or better discriminative performance. From a clinical perspective, this reduction is critical: fewer alarms mean less alarm fatigue, fewer interruptions, and a higher likelihood that each alert receives appropriate attention. The model therefore optimizes not only standard statistical metrics but also a key operational constraint in intensive care.

4.4. Alert lead time

Our system generates early warnings before the index event (e.g., sepsis, acute kidney injury). For each event, we compute the lead-time between the first high-risk alert and the occurrence of the clinical outcome, and summarize it via the median and interquartile range (IQR). A longer lead-time, under a controlled alert burden, increases the clinical usefulness of the system by allowing proactive interventions instead of late, confirmatory alarms. In our experiments, the model consistently produces meaningful lead times before critical events, demonstrating its potential to support anticipatory clinical decision-making rather than merely flagging already obvious deteriorations.

4.5. Fairness across subgroups

We evaluated performance across key clinical subgroups to assess fairness and stability: age (<65 vs. ≥65 years) and diabetes status (diabetic vs. non-diabetic). Table 4 summarizes the results.

Performance remains high and stable across all groups, with AUC values between 0.88, 0.90, and F1-scores between 0.60 and 0.64. Differences in false positive and false negative rates are small and non-systematic, with $|\Delta FPR|$ and $|\Delta FNR| \leq 0.02$. These findings indicate that no subgroup is disproportionately disadvantaged and that the model maintains an approximately balanced error profile across populations. If a clinically unjustified discrepancy were to emerge in practice, the pipeline can apply context-specific adjustments and subgroup calibration to further mitigate potential biases and support fair use in heterogeneous ICU populations.

Table 4. Fairness across subgroups

Group	Prevalence	AUC	F1	ΔFPR	ΔFNR
<65 years	0.52	0.90	0.64	-	-
≥65 years	0.48	0.88	0.60	+0.01	+0.02
Diabetic	0.22	0.89	0.63	+0.00	+0.01
Not diabetic	0.78	0.89	0.62	+0.00	+0.01

4.6. Concrete examples and explanations (XAI)

For each alert, the system generates a concise explanation report that includes:

- The top-k contributing variables, extracted using SHAP/LIME;

- The most influential time windows in the recent patient trajectory;
- The activated context rules that influenced the final anomaly score or alert threshold.

Clinician feedback indicates that these explanations improve the understandability and perceived credibility of alerts and help with prioritization. Although XAI does not significantly increase AUC or F1, it is transformative for clinical adoption and auditing: this design transforms the model from a black-box scoring engine into a transparent decision-support tool that clinicians can inspect, discuss, and refine.

4.7. Robustness and sensitivity

We performed a series of stress tests to assess the robustness of the model to realistic perturbations, including:

- Simulated missing data,
- Added noise to input variables,
- Temporal shifts in input sequences.

Overall, the model maintained stable performance under these perturbations, with variations generally within a few percentage points of baseline metrics. In scenarios where we observed localized degradation, we applied targeted retraining and rule adjustments to restore performance or local recalibration was sufficient to restore performance. These results show that we can maintain and adapt the system over time as clinical practice, data collection patterns, and patient populations evolve.

4.8. Portability

We also evaluated the portability of the model across different times and ICU wards within the MIMIC database. Performance remained consistent across units, with any differences largely attributable to variations in cohort composition rather than model instability. This stability across settings suggests that the model captures robust physiological patterns rather than overfitting to a single ward or time window. It supports the feasibility of scaling the approach to other units or institutions, subject to appropriate local validation and calibration [25].

4.9. Discussion

Our results show that the proposed hybrid-alerting framework does more than provide incremental gains over existing approaches. Compared with clinical thresholds and strong deep learning baselines, it achieves substantial improvements in ROC AUC, PR AUC, and F1-score, while reducing the overall alert burden by roughly one third. For clinicians, this combination-fewer alerts but with higher relevance, directly addresses alarm fatigue while preserving early detection capability.

A key contribution of our work is the emphasis on probability calibration. Many high-performing models in critical care output scores that are difficult to interpret as risks. By explicitly calibrating predicted probabilities, we obtain risk estimates that better match observed event frequencies and are easier to use for threshold selection, resource allocation, and communication of risk. This makes it more natural to integrate the model into existing decision-making processes.

The framework also maintains relatively stable and fair performance across major patient subgroups and under robustness checks. Although we cannot exclude all sources of bias, the combination of contextualization, calibration, and monitoring helps limit performance disparities. In particular, the contextual layer adjusts anomaly scores based on chronic conditions and treatment profiles, reducing obvious false positives when abnormal values are clinically expected.

We integrate XAI as a core component rather than treating it as an afterthought. For each alert, the system provides concise explanations highlighting key variables and influential time windows. This transforms the model from a black-box scoring engine into a transparent decision-support tool that clinicians can inspect, discuss, and refine, improving trust, auditability, and the ability to reconcile model output with clinical intuition.

Finally, we embed the model in a governance-oriented pipeline that includes drift monitoring, fairness checks, logging, and recalibration procedures, and we align our reporting with emerging clinical AI standards. In the broader landscape of decision support, our approach sits between purely black-box models and simple rule-based systems: it combines sequential deep learning with explicit patient context, integrated explainability, and calibrated outputs, offering a practical foundation for AI-based alerting in intensive care.

4.9.1. Runtime and computational complexity

Although the proposed framework combines several components (sequential backbone, contextual layer, XAI, and calibration), its computational cost remains compatible with near real-time deployment in an ICU setting. The core sequential models (autoencoders and Transformer) have a runtime that scales approximately linearly with the length of the input sequence and the number of variables (O (T·d)), and the

contextual adjustments are implemented as lightweight feed-forward operations on precomputed patient profiles.

We perform probability calibration once per model after training, so it does not affect inference latency. We compute explainability methods (SHAP, LIME, and attention visualizations) only for triggered alerts and run them asynchronously or at a lower frequency, which limits their impact on overall throughput. In practice, this design allows the system to update risk scores and alerts frequently enough to support continuous monitoring without imposing prohibitive computational requirements.

4.9.2. Limitations

Despite these promising results, this study has several limitations. First, we base our analysis on retrospective data from a single ICU database (MIMIC-III/IV) and conduct all experiments offline, so we have not yet evaluated real-time performance or its impact on clinical workflows and patient outcomes. Second, we primarily evaluate the model on a limited set of deterioration events (e.g., sepsis, acute kidney injury), and we do not yet know how well the same architecture, thresholds, and contextual rules generalize to other endpoints or care settings. Third, although we assess fairness across age and diabetes status, we still need a broader analysis across additional demographic and clinical dimensions to better characterize potential biases.

We also tune the contextual rules and calibration procedures on the same overall data source, so we may need to adapt, re-validate, and recalibrate the system when we deploy it in new institutions with different case mix, documentation practices, or monitoring protocols. In addition, we do not perform extensive formal hypothesis testing across all model comparisons (e.g., for all differences in AUC or F1), and statistical analyses are limited to bootstrap-based confidence interval estimation for selected key comparisons, so readers should interpret performance differences as indicative rather than statistically confirmed. Finally, we do not carry out a detailed case-level error analysis of typical false positives and false negatives with clinicians, nor a systematic benchmarking of runtime performance or computational cost; we leave both a qualitative error analysis and a more detailed study of latency, scalability, and resource usage for future work.

4.9.3. Future work

Future work could extend this approach in several directions. First, scaling up personalization by integrating richer longitudinal trajectories and comorbidity profiles could enable finer risk stratification and more tailored alerts. Second, federated learning could help preserve data privacy while enabling training and adaptation across multiple institutions. Third, human-AI interaction studies should investigate how clinicians interpret and act on AI-generated explanations in real time. Fourth, continuous validation pipelines with real-world feedback loops could dynamically adjust thresholds, rules, and calibration as practice and populations evolve.

Finally, more systematic evaluations of ethics and fairness will be essential to ensure equitable behaviour across patient subgroups. In summary, our hybrid approach not only improves the clinical relevance and interpretability of alerts but also lays the groundwork for trustworthy, continuously monitored, and governance-aligned AI systems in healthcare.

5. CONCLUSION

In this paper, we proposed a hybrid-alerting framework that combines sequential modeling, patient context, XAI, and probability calibration to support early detection of clinical deterioration in intensive care. Compared with clinical thresholds and deep learning baselines, the proposed model achieved higher discrimination, improved precision–recall performance, and better calibration, while reducing the alert rate by approximately one third. These improvements translate into fewer but more informative alerts, improved risk stratification at the bedside, and a lower risk of alarm fatigue for clinical staff.

The work makes three main contributions.

First, it shows that integrating sequential models with contextual information, explanation mechanisms, and probability calibration within a single architecture can simultaneously improve discrimination, calibration, and alert burden relative to both clinical thresholds and strong deep learning baselines. Second, it demonstrates that contextualization and calibration can reduce alarm fatigue while preserving early detection and stable performance across key patient subgroups, yielding risk scores that are both numerically reliable and operationally useful. Third, it embeds the modeling approach in a governance-oriented pipeline that includes drift monitoring, fairness checks, alert logging, and alignment with emerging clinical AI reporting practices, positioning the system as a realistic candidate for deployment in critical care environments.

Beyond predictive performance, we designed the system with real-world deployment in mind. For each alert, we provide an interpretable explanation. We monitor model behavior over time and integrate maintenance procedures such as recalibration and rule updates into the pipeline. Overall, the proposed framework provides a concrete pathway toward trustworthy and monitored AI-based alerting in critical care

and forms a basis for future multi-centre validation, large-scale personalization, and prospective evaluation of human-AI collaboration at the bedside.

ACKNOWLEDGMENTS

The authors would like to thank their laboratory and institution for their support during this research.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Amel Elketroussi	✓	✓	✓						✓	✓				
Bachir Djebbar				✓	✓					✓				
Ibtissem Bekkouche		✓		✓						✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY

The data used in this study are publicly available from the MIMIC-III and MIMIC-IV databases (PhysioNet).

REFERENCES

[1] Y. Su, R. Liu, Y. Zhao, W. Sun, C. Niu, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Jul. 2019, pp. 2828–2837. doi: 10.1145/3292500.3330672.

[2] Y. Deng, S. Liu, Z. Wang, Y. Wang, Y. Jiang, and B. Liu, "Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients," *Frontiers in Medicine*, vol. 9, Sep. 2022, doi: 10.3389/fmed.2022.933037.

[3] M. Moutaib, M. Fattah, Y. Farhaoui, B. Aghoutane, and M. El Bekkali, "Fetal electrocardiogram prediction using machine learning: a random forest-based approach," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 33, no. 2, pp. 1076-1083, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp1076-1083.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?,'" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, NY, USA: ACM, Aug. 2016, pp. 1135-1144. doi: 10.1145/2939672.2939778.

[5] Lundberg Scott M. and Lee Su-In., " A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*., 2017, pp. 4765-4774.

[6] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[7] M. Saarela and V. Podgorelec, "Recent applications of explainable AI (XAI): a systematic literature review," *Applied Sciences*, vol. 14, no. 19, p. 8884, Oct. 2024, doi: 10.3390/app14198884.

[8] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. suppl_1, 1 January 2004, pp. D267-D270, Jan. 2004, doi: 10.1093/nar/gkh061.

[9] SNOMED International, "SNOMED CT clinical terms – release," 2023. [Online]. Available: https://www.snomed.org.

[10] Health Level Seven International, "HL7 FHIR (Fast Healthcare Interoperability Resources)," 2019. [Online]. Available: https://www.hl7.org/fhir.

[11] B. J. Drew *et al.*, "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients," *PLoS ONE*, vol. 9, no. 10, p. e110274, Oct. 2014, doi: 10.1371/journal.pone.0110274.

[12] R. C. of Physicians, "National early warning score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS," London, 2017. [Online]. Available: https://www.rcp.ac.uk/media/a4ibkkbf/news2-final-report_0_0.pdf

[13] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: A severity of disease classification system," *Critical Care Medicine*, vol. 13, no. 10, pp. 818-829, Oct. 1985, doi: 10.1097/00003246-198510000-00009.
[14] E. Choi, M. Taha Bahadori, J. Sun, J. A. Kulas, A. Schuetz, and W. F. Stewart, "Retain: an interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html
[15] Y. Li *et al.*, "BEHRT: transformer for electronic health records," *Scientific Reports*, vol. 10, no. 1, p. 7155, Apr. 2020, doi: 10.1038/s41598-020-62922-y.
[16] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digital Medicine*, vol. 4, no. 1, p. 86, 2021, doi: 10.1038/s41746-021-00455-y.
[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, 2017, pp. 1321-1330.
[18] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447-453, Oct. 2019, doi: 10.1126/science.aax2342.
[19] Q. Feng, M. Du, N. Zou, and X. Hu, "Fair machine learning in healthcare: a survey," *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 3, pp. 493–507, Mar. 2025, doi: 10.1109/TAI.2024.3361836.
[20] K. Chakradeo *et al.*, "Navigating fairness aspects of clinical prediction models," *BMC Medicine*, vol. 23, no. 1, p. 567, Oct. 2025, doi: 10.1186/s12916-025-04340-3.
[21] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.
[22] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV v2.0," *PhysioNet*, 2022, doi: 10.13026/7vcr-e114.
[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[24] R. E. Ko *et al.*, "Deep learning-based early warning score for predicting clinical deterioration in general ward cancer patients," *Cancers*, vol. 15, no. 21, p. 5145, Oct. 2023, doi: 10.3390/cancers15215145.
[25] B. Vasey *et al.*, "Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Bmj*, vol. 377, p. e070904, May 2023, doi: 10.1136/bmj-2022-070904.

BIOGRAPHIES OF AUTHORS

Amel Elketroussi is a lecturer in the Department of Computer Science, Faculty of Mathematics and Computer Science, University of Science and Technology of Oran – Mohamed Boudiaf (USTO-MB), Algeria. She obtained her Magister degree in computer science from the same university. Her research interests focus on databases and artificial intelligence (AI), with particular emphasis on data management and intelligent information systems. She is actively involved in academic research projects and contributes to collaborative research activities within her department. Her teaching responsibilities include database systems, fundamentals of machine learning, and AI-based applications in data processing, reflecting her commitment to both education and research in computer science. She can be contacted at email: amel.elketroussi@univ-usto.dz.

Bachir Djebbar doctoral degree. in mathematical sciences from Paul Sabatier University in Toulouse, France in 1987, and Ph.D. State in mathematical approximation from the University of Science and Technology of Oran. He held several administrative positions at the Faculty of Mathematics and Computer Science at the Mohamed Boudiaf University of Technology in Oran (USTOMB). Head of the Computer Science Department from 1999 to 2008, Vice Dean in charge of studies from 2008 to 2011 and Dean of the Faculty of Mathematics and Computer Science, since 2017, he is currently a professor in the Department of Computer Science. He can be contacted at email: bachir.djebbar@univ-usto.dz.

Ibtissem Bekkouche is an Associate Professor in the ISD team at the SIMPA Laboratory, Faculty of Mathematics and Computer Science, University of Science and Technology of Oran – Mohamed Boudiaf (USTO-MB), Algeria. She earned her Ph.D. in computer science from the same university, specializing in pattern recognition and artificial intelligence. Her research focuses on artificial intelligence, machine learning, computer vision, image processing, and data science. She can be contacted at email: ibtissem.bekkouche@univ-usto.dz.