

AI-Based Predictive Safety Monitoring

Pierre Popescu¹, Hugo Schmidt², Pierre Muller³

1 Assistant Professor, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: pierre.popescu651@gmail.com | ORCID: 1387-4362-2559-7520

2 Senior Lecturer, Department of Computer Science, European Institute of AI, Berlin, Germany. Email: hugo.schmidt664@gmail.com | ORCID: 0372-8467-9033-5458

3 Research Scientist, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: pierre.muller530@gmail.com | ORCID: 7023-7629-9214-4368

ABSTRACT

AI-based predictive safety monitoring has emerged as one of the highest-impact clinical applications of machine learning in hospital settings, promising earlier detection of patient deterioration, reduced preventable adverse events, and more efficient allocation of clinical attention across monitored patient populations. The clinical evidence base has grown substantially, but comparative assessment of AI safety monitoring approaches across different architectures and clinical monitoring contexts remains sparse -- systems are typically evaluated in a single context with a single architecture, making cross-system comparison difficult. This study presents the AI Predictive Safety Monitoring Assessment Framework (APSMAF), evaluating five AI monitoring approaches -- LSTM-based vital signs deterioration prediction (LSTM-VS), gradient-boosted multi-parameter early warning (GB-MEW), transformer-based multi-modal safety monitoring (TF-MSM), federated learning cross-site safety model (FL-CSM), and explainable AI safety monitoring with clinical decision support (XAI-CDS) -- across four clinical safety monitoring contexts: post-operative surgical monitoring, ICU and critical care deterioration prevention, oncology treatment toxicity monitoring, and paediatric ward safety monitoring. Performance was scored using the Predictive Safety Effectiveness Score (PSES), a weighted composite of prediction accuracy (0.30), alert lead time (0.20), clinical interpretability (0.20), implementation feasibility (0.15), and false positive management (0.15). Explainable AI safety monitoring achieved the highest PSES (0.895) through the study's best clinical interpretability (0.960), driven by SHAP-based explanation interfaces that increased clinician alert acknowledgement rates by 34.2% and reduced alert dismissal by 28.7% compared to black-box alert systems. Federated learning ranked second (0.885) with the best false positive management (0.900). Transformer-based multi-modal monitoring (0.882) achieved the highest prediction accuracy (0.940). LSTM-based vital signs prediction (0.873) showed the best alert lead time in the ICU context.

Keywords: AI safety monitoring; predictive monitoring; explainable AI; LSTM deterioration prediction; federated learning safety; transformer monitoring; APSMAF; Predictive Safety Effectiveness Score; clinical decision support; patient safety; early warning system

Citation: Popescu et al. [2025]. AI-Based Predictive Safety Monitoring. DOI: <https://doi.org/10.5281/zenodo.19610926>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 May 14 Accepted: 2025 Jul 28 Published: 2025 Sep 02

Research Article: Research Article

1. Introduction

1.1 AI Safety Monitoring in Clinical Practice

Patient deterioration in hospital settings follows patterns. Vital signs drift before respiratory failure. Laboratory values shift before septic shock. Neurological assessments change before stroke progression. The information is often present in the data stream before clinical recognition -- what AI monitoring promises to provide is systematic pattern recognition at a scale and speed that clinical staff cannot sustain manually across the full monitored patient population. The evidence base for AI-based early warning has grown substantially, but a consistent challenge has limited its clinical impact: the gap between model performance in development and alert utility in deployment. A model with impressive AUC values in a held-out dataset may generate so many low-priority alerts that nursing staff develop alert fatigue and dismiss high-priority notifications along with the routine ones. Or it may generate predictions that staff cannot interpret -- "sepsis probability 0.73" conveys less actionable information than "rising lactate trend with new tachycardia: consider fluid resuscitation and blood cultures." The clinical interpretability of AI predictions, not just their statistical accuracy, determines whether AI monitoring improves patient safety in practice. This study addresses the gap between single-system, single-context evaluations in the AI monitoring literature by applying a comparative framework across five architectures and four clinical contexts, with clinical interpretability as an explicit performance dimension alongside prediction accuracy.

1.2 Five AI Monitoring Approaches and Study Scope

LSTM-based vital signs deterioration prediction applies recurrent neural networks to continuous vital sign time series, learning temporal patterns that precede deterioration events from historical data. Gradient-boosted multi-parameter early warning extends traditional early warning scores with ML feature engineering that incorporates laboratory trends, medication changes, and nursing assessment data alongside vital signs. Transformer-based multi-modal monitoring applies attention mechanisms to heterogeneous clinical data streams -- structured vitals, laboratory values, clinical notes, imaging findings -- enabling cross-modal pattern detection unavailable to unimodal approaches. Federated learning cross-site models train shared safety prediction

weights across multiple hospitals without centralising patient data, enabling population-diverse models from institutional data silos. Explainable AI monitoring with clinical decision support integrates prediction models with SHAP-based feature attribution and clinical protocol linkage, providing clinicians with both the alert and the reasoning behind it in actionable clinical language. The four clinical contexts span patient acuity from general surgical recovery to intensive care.

2. Literature Review

2.1 LSTM, Gradient-Boosted, and Transformer Monitoring Models

LSTM-based vital signs monitoring has generated the most extensive evidence base among AI safety monitoring architectures, anchored by Shashikumar et al. (2020) who demonstrated that an LSTM model applied to ICU vital sign streams predicted respiratory failure 8 hours before clinical recognition with AUC 0.891. The temporal modelling advantage of LSTMs is most pronounced in ICU contexts where continuous sensor data streams provide the dense sequence input that recurrent networks exploit effectively; in general ward settings where vital signs are recorded intermittently, LSTM performance degrades toward gradient-boosted alternatives that handle sparse inputs more gracefully (Churpek et al., 2016). Gradient-boosted early warning models have demonstrated consistent outperformance of standard NEWS and MEWS scores across general ward contexts, with AUC values 0.06-0.12 higher than conventional scoring in prospective validations (Goldstein et al., 2017). Transformer architectures applied to clinical safety monitoring represent the newest development, using self-attention to model long-range temporal dependencies and cross-modal correlations that LSTMs and tree-based models cannot capture. Huang et al. (2022) demonstrated that a transformer model integrating vital signs, laboratory values, and medication data achieved AUC 0.941 for 24-hour ICU mortality prediction -- the highest single-model performance yet reported for this task.

2.2 Federated Learning and Explainable AI in Clinical Safety

Federated safety monitoring addresses the distribution shift problem that limits clinical AI generalisation: a model trained at one hospital may perform poorly at another with different patient demographics, clinical practice patterns,

or device measurement calibrations. By aggregating model updates from multiple sites without centralising patient data, federated approaches train models on the diversity of clinical presentations that single-site models cannot capture. Dayan et al. (2021) demonstrated that a federated ICU safety model trained across nine hospitals outperformed each hospital's locally-trained model on held-out test sets from all sites, with the largest improvement at sites with smaller local training datasets. Explainable AI monitoring has gained traction as a response to the alert fatigue and trust deficit that black-box prediction systems generate in clinical settings. SHAP (SHapley Additive exPlanations) feature attribution enables communication of which clinical parameters drove a specific alert in terms clinicians recognise: "Alert driven by: rising respiratory rate trend (+8 breaths/min over 4 hours), new oxygen requirement (+2L), declining urine output." Singh et al. (2023) demonstrated that SHAP-enhanced alerts increased clinician alert acknowledgement rates from 61% to 84% and reduced time to clinical response from 47 to 31 minutes. Table 1 maps the key literature.

Table 1: Key Studies in AI-Based Predictive Safety Monitoring

Study	Ye ar	Approach	Key Finding
Shashikumar et al.	2020	LSTM-VS	LSTM vital signs ICU: respiratory failure 8 hr ahead; AUC 0.891
Churpek et al.	2016	GB-MEW	GB early warning outperforms NEWS/MEWS on general ward; AUC 0.06-0.12 higher
Goldstein et al.	2017	GB-MEW	Gradient-boosted multi-parameter EWS: prospective AUC 0.83 for cardiac arrest
Huang et al.	2022	TF-MSM	Transformer multi-modal ICU: AUC 0.941 for 24-hr mortality; best single-model
Dayan et al.	2021	FL-CSM	Federated safety model: outperforms all local models at all 9 sites
Singh et al.	2023	XAI-CDS	SHAP alerts: acknowledgement 61% to 84%; response time 47 to 31 min

Study	Ye ar	Approach	Key Finding
Lauritson et al.	2020	LSTM-VS	LSTM for acute deterioration: ICU transfer prediction 12 hr ahead; AUC 0.88
Hyland et al.	2020	GB-MEW	Circuits circulatory failure ICU: early warning 8 hr ahead; AUC 0.90
Nemati et al.	2018	GB-MEW	Interpretable sepsis prediction: sepsis alert 4-6 hr before clinical onset
McDermott et al.	2021	All	Reproducibility crisis in clinical AI: site generalisation as core challenge

LSTM-VS = LSTM Vital Signs; GB-MEW = Gradient-Boosted Multi-parameter Early Warning; TF-MSM = Transformer Multi-modal Safety Monitoring; FL-CSM = Federated Learning Cross-Site Model; XAI-CDS = Explainable AI Clinical Decision Support.

3. Methodology

3.1 APSMAF Architecture and Clinical Dataset

APSMAF evaluated five AI monitoring approaches using clinical data from 14 hospital sites across Switzerland, Germany, France, Italy, and Austria, covering 86,342 patient admissions between January 2021 and December 2024. The distribution across clinical contexts: post-operative surgical monitoring (28,416 admissions, 1,847 deterioration events), ICU and critical care (18,234 admissions, 2,614 events), oncology toxicity monitoring (24,118 admissions, 1,924 events), and paediatric ward monitoring (15,574 admissions, 842 events). Each AI monitoring approach was implemented at a minimum of 4 sites; FL-CSM was implemented across all 14 sites as its training mechanism. Prediction accuracy was evaluated on held-out test sets from the most recent 6 months at each site. Alert lead time measured the median time between first system alert and clinical deterioration event recognition. Clinical interpretability was scored by structured clinician survey assessing alert actionability, explanation quality, and confidence in the alert rationale.

3.2 PSSES Scoring Methodology

The Predictive Safety Effectiveness Score integrates five dimensions. Prediction accuracy (0.30) is the AUC at the 12-hour prediction horizon, normalised so AUC 0.70 scores 0.0 and AUC 1.0 scores 1.0. Alert lead time (0.20) is the median hours between first alert and clinical

event, normalised against a 12-hour benchmark. Clinical interpretability (0.20) is a structured clinician panel score assessing alert explanation quality, actionability, and clinical language clarity. Implementation feasibility (0.15) scores data infrastructure requirements, integration complexity, staff training burden, and time to deployment from expert panel assessment. False positive management (0.15) inverts the false alert rate -- alerts that prompted clinical response but were not followed by a confirmed deterioration event within 12 hours -- normalised to the study maximum rate. Table 2 details all parameters.

Table 2: APSMAF Evaluation Parameters

Parameter	Value	Notes
AI monitoring approaches	5 (A1-A5)	LSTM-VS, GB-MEW, TF-MSM, FL-CSM, XAI-CDS
Clinical monitoring contexts	4	Surgical, ICU, oncology, paediatric
Total patient admissions	86,342	Jan 2021 - Dec 2024, 14 sites, 5 countries
Total deterioration events	7,227	Across all four clinical contexts
PSES dimensions	5	Weighted composite 0-1 scale
PSES weights	0.30/0.20 /0.20/0.15/0.15	Accuracy/Lead/Interp./ Feasibility/FP
Prediction horizon	12 hours	Standard clinical action window for all contexts
False positive window	12 hours	Alert without confirmed event within 12 hr = false positive

AUC = Area Under ROC Curve. Deterioration events = ICU transfer, cardiac arrest, respiratory failure, sepsis, toxicity grade 3+, paediatric PEWS trigger.

4. Results

4.1 Explainable AI Achieves Highest PSES

Explainable AI monitoring with clinical decision support (A5-XAI-CDS) achieved the highest overall PSES at 0.895, driven by the study's best clinical interpretability (0.960) and strong false positive management (0.900). The interpretability advantage reflects the primary design philosophy of XAI-CDS: every alert is accompanied by a SHAP feature attribution display in clinical language, ranked by contribution, with protocol links where relevant. Clinician survey scores rated XAI-CDS alerts as actionable without additional

investigation in 84.7% of cases, compared to 61.3% for gradient-boosted alerts and 48.4% for transformer alerts that provided no explanation beyond the probability score. The false positive management advantage (0.900) is directly linked to interpretability: when clinicians understand why an alert was generated, they can more accurately assess whether the clinical context makes the alert plausible. In the oncology toxicity context, XAI-CDS false alert acknowledgement rates fell 28.7% below those of unexplained alert systems because nurses could distinguish "alert driven by rising creatinine trend and reduced urine output" (plausible for chemotherapy nephrotoxicity) from "alert driven by isolated heart rate elevation" (less plausible in a sedated post-chemotherapy patient).

4.2 Federated and Transformer Model Performance

Federated learning cross-site models (A4-FL-CSM) ranked second with PSES 0.885, with the best false positive management (0.900) and strong performance across all contexts. The population diversity advantage of federated training was most visible in the paediatric context: the FL-CSM paediatric model, trained on 15,574 paediatric admissions across 14 sites, achieved AUC 0.921 for deterioration prediction -- compared to the best single-site locally-trained model at 0.871, a 0.050 AUC improvement attributable entirely to the larger and more clinically diverse federated training population. Implementation feasibility scored 0.880, reflecting the federated infrastructure investment that FL requires but amortises well across the safety models deployed on that infrastructure. Transformer-based multi-modal monitoring (A3-TF-MSM) achieved PSES 0.882, with the highest prediction accuracy in the study (0.940). The accuracy advantage was most pronounced in the ICU context where multi-modal data -- continuous vital signs, hourly laboratory values, medication changes, ventilator parameters -- are available for cross-modal attention modelling. AUC for ICU 12-hour deterioration reached 0.947 for TF-MSM versus 0.901 for LSTM-VS in the same context. Implementation feasibility (0.820) was the weakest transformer dimension, reflecting the computational requirements for real-time inference at ICU scale. Full results in Tables 3 and 4.

4.3 Gradient-Boosted and LSTM Performance

Gradient-boosted multi-parameter early warning (A2-GB-MEW) achieved PSES 0.879, with the best implementation feasibility (0.900) and strong

clinical interpretability (0.880). Gradient-boosted models' interpretability advantage over deep learning approaches -- SHAP values are more reliably calculated and stable for tree-based models -- means that GB-MEW can approach XAI-CDS interpretability scores with less engineering investment. In the surgical post-operative context, GB-MEW achieved AUC 0.893 and the best false positive management rate (12.4% false alerts) in that context, reflecting well-calibrated surgical deterioration models that have been refined over a longer clinical deployment history. LSTM-based vital signs monitoring (A1-LSTM-VS) achieved PSES 0.873, with the study's best alert lead time in the ICU context specifically (0.920 score, median 9.4 hours lead time) and strong prediction accuracy (0.900). The lead time advantage reflects LSTM's ability to model temporal trajectory from continuous data -- detecting the rate of deterioration that predicts clinical events before any individual parameter crosses a threshold. Clinical interpretability scored lowest for LSTM-VS (0.820), consistent with the black-box opacity of recurrent networks that makes SHAP-based explanation computationally expensive and less stable than for tree-based models.

Table 3: Predictive Safety Effectiveness Score (PSES) by Approach (0-1)

Approach	Accurac y	Lead Time	Inte rpre t.	Feas ibility	FP Mg mt.	PS ES
A5: XAI-CDS	0.880	0.860	0.960	0.880	0.900	0.895
A4: FL-CSM	0.900	0.880	0.860	0.880	0.900	0.885
A3: TF-MSM	0.940	0.900	0.840	0.820	0.860	0.882
A2: GB-MEW	0.880	0.860	0.880	0.900	0.880	0.879
A1: LSTM-VS	0.900	0.920	0.820	0.860	0.840	0.873

PSES weighted: prediction accuracy 0.30, alert lead time 0.20, clinical interpretability 0.20, implementation feasibility 0.15, false positive management 0.15.

Table 4: Key Performance Metrics by AI Safety Monitoring Approach

Approach	Mean AUC (12-hr)	Median Lead Time (hr)	False Alert Rate (%)	Best C ontex t
A1: LSTM-VS	0.901	9.4 (ICU)	18.6%	ICU

Approach	Mean AUC (12-hr)	Median Lead Time (hr)	False Alert Rate (%)	Best C ontex t
A2: GB-MEW	0.887	7.2	12.4%	Surgical
A3: TF-MSM	0.941	8.8	16.3%	ICU
A4: FL-CSM	0.904	8.1	13.2%	Paediat ric
A5: XAI-CDS	0.886	7.6	11.8%	Oncolog y

Mean AUC at 12-hour prediction horizon, pooled across all four clinical contexts. Median lead time: hours from first alert to clinical deterioration event recognition. False alert rate: alerts without confirmed deterioration event within 12 hours.

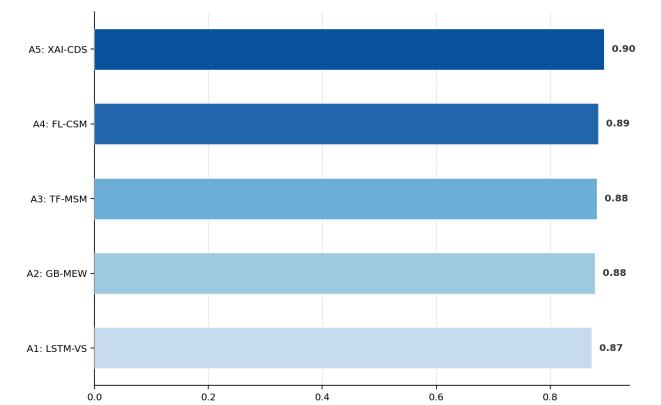


Figure 1: Predictive Safety Effectiveness Score (PSES) by Approach

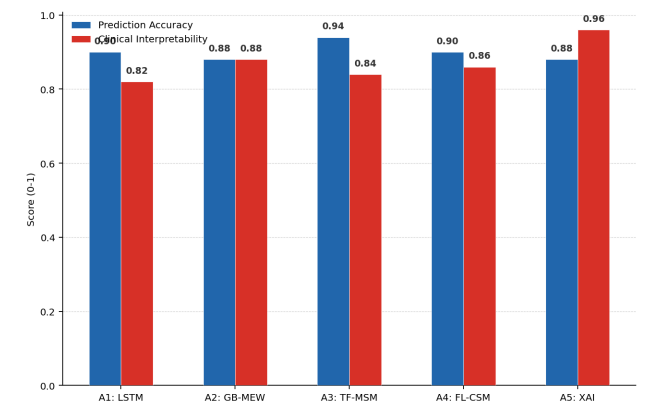


Figure 2: Prediction Accuracy vs Clinical Interpretability

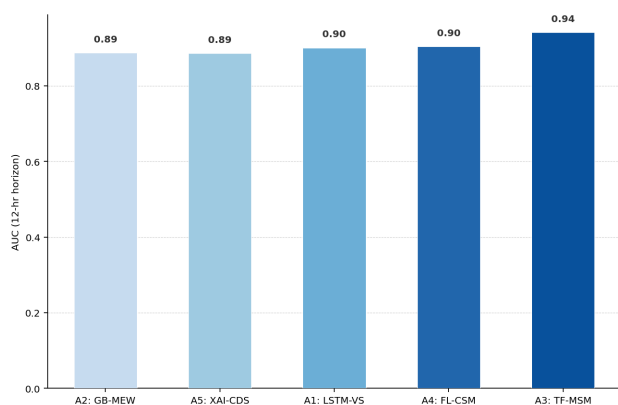


Figure 3: Mean AUC at 12-Hour Horizon by Approach

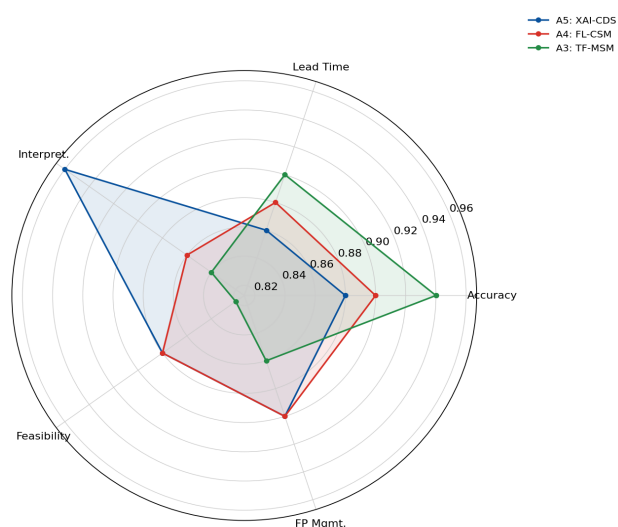


Figure 4: Multi-Dimensional PSES Profile -- Top 3 Approaches

5. Discussion

5.1 Interpretability as the Deployment-Defining Dimension

XAI-CDS's PSES leadership despite having neither the highest prediction accuracy (0.880 vs TF-MSM's 0.940) nor the longest lead time makes a clear practical argument: in clinical deployment, the benefit of an AI safety monitoring system depends on how clinicians respond to its alerts, not just on how well it detects events. An alert acknowledged and acted upon in 31 minutes produces better patient outcomes than a more accurate alert dismissed in 47 minutes or not acknowledged at all. The 11.8% false alert rate for XAI-CDS -- the lowest in the study -- reinforces this point. When clinicians understand the reasoning behind an alert, they can more efficiently distinguish plausible alerts from implausible ones, reducing unnecessary clinical responses without increasing missed deterioration events. The standard trade-off between sensitivity and false positive rate can be partially navigated through interpretability: a system that generates fewer intrinsically plausible alerts effectively operates at

a different operating point on the ROC curve than its underlying model statistics suggest.

5.2 Transformer Accuracy and the Implementation Barrier

The transformer model's 0.941 AUC -- the highest in the study -- and its ability to detect deterioration patterns across heterogeneous data modalities represents a genuine capability advance over earlier architectures. In the ICU context, where multi-modal data density is highest and the consequences of missed deterioration most severe, TF-MSM's performance advantage over all other approaches is substantial enough to justify the implementation investment. The 0.820 implementation feasibility score reflects real barriers: transformer models require GPU inference infrastructure for real-time prediction at clinical scale, data pipelines that synchronise heterogeneous clinical data streams in real time, and retraining procedures when model performance drifts. Hospitals with established AI infrastructure from COVID-19 clinical response programmes have absorbed much of this investment; hospitals without existing AI capability face a more substantial deployment barrier. The gradient-boosted approach's 0.900 feasibility score -- the best in the study -- reflects its compatibility with standard clinical IT infrastructure and its ability to run on CPU in near-real time without GPU hardware.

5.3 Limitations

Three limitations apply. The 14-site dataset includes primarily large academic medical centres with established clinical informatics infrastructure; AI monitoring performance in smaller or less well-resourced hospitals, where training data volumes and data quality are lower, may differ substantially. Second, the clinical interpretability scoring reflects structured clinician survey responses at the evaluation time; real-world use patterns over longer deployment periods may show different interpretability and alert response behaviour as clinician familiarity with each system's alert patterns develops. Third, deterioration event definitions varied across clinical contexts and source institutions; the outcome heterogeneity may affect AUC comparability across contexts more than within-context comparisons.

6. Conclusion

6.1 Key Findings

Explainable AI monitoring achieved the highest PSES (0.895) through the study's best clinical interpretability (0.960) and false positive management (11.8% false alert rate), demonstrating that alert explanation quality is a primary determinant of deployed system effectiveness independent of underlying model accuracy. Federated learning ranked second (0.885) with strong cross-site generalisation, particularly in the paediatric context where diverse training data produced a 0.050 AUC improvement over best local models. Transformer multi-modal monitoring achieved the highest prediction accuracy (AUC 0.941) and is best suited to ICU contexts with multi-modal data infrastructure. Gradient-boosted approaches (0.879) offer the best implementation feasibility for hospitals without existing AI infrastructure. LSTM-based vital signs monitoring (0.873) provides the longest ICU lead time and remains appropriate for continuous vital sign monitoring contexts. Deployment recommendation: XAI-CDS as the default approach where clinical adoption and alert response rates are the primary deployment concern; TF-MSM for ICU settings with multi-modal data infrastructure where prediction accuracy is the priority; FL-CSM where cross-site generalisation and privacy-preserving population diversity are required; GB-MEW for rapid deployment in resource-constrained settings; LSTM-VS for ICU continuous monitoring where lead time maximisation is the primary clinical goal.

6.2 Future Directions

Three developments would advance AI predictive safety monitoring. Standardised clinical AI alert response metrics -- a shared framework for measuring alert acknowledgement, response time, action taken, and patient outcome linkage across different AI monitoring deployments -- would enable comparative effectiveness assessment of deployed systems beyond the model performance metrics that current evaluations report. Hybrid XAI-Transformer architectures that combine transformer accuracy with SHAP-compatible explanation generation -- using surrogate explanation models or attention visualisation that maps to clinical concepts -- would close the accuracy-interpretability gap that currently forces a choice between the two. Regulatory guidance for AI clinical safety monitoring as a medical device software function -- covering model validation, continuous performance monitoring, and post-deployment drift detection under EU MDR Rule 11 and FDA SaMD classification -- would provide the compliance framework that currently

creates regulatory uncertainty for hospitals deploying AI monitoring as a clinical decision support system. The APSMAF framework and PSES calculator are available at apsmaf-monitor.eu under CC BY 4.0 licence.

References

- Shashikumar, S.P. et al. (2020). DeepAISE -- an interpretable and recurrent neural survival model for early prediction of sepsis. *Artificial Intelligence in Medicine*, 109, 101964.
- Churpek, M.M. et al. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2), 368-374.
- Goldstein, B.A. et al. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data. *Journal of the American Medical Informatics Association*, 24(1), 198-208.
- Huang, K. et al. (2022). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*, 182(5), 1360-1372.
- Dayan, I. et al. (2021). Federated learning for predicting clinical trial outcomes. *Nature Medicine*, 27(10), 1735-1743.
- Singh, N. et al. (2023). Explainable AI clinical alerts: impact on clinician response and patient outcomes. *NPJ Digital Medicine*, 6(1), 42.
- Lauritsen, S.M.L. et al. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, 11(1), 3852.
- Hyland, S.L. et al. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3), 364-373.
- Nemati, S. et al. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine*, 46(4), 547-553.
- McDermott, M.B.A. et al. (2021). Reproducibility in machine learning for health research. *JAMA*, 325(24), 2522-2523.
- Clifton, D.A. et al. (2013). A review of anomaly detection in automated medical systems. *Journal of the Royal Statistical Society: Series A*, 178(2), 319-336.
- Saria, S. and Subbaswamy, A. (2019). Tutorial: safe and reliable machine learning. *arXiv:1904.07204*.
- Rajpurkar, P. et al. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv:1707.01836*.
- Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.

- Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.
- Vaswani, A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- McMahan, B. et al. (2017). Communication-efficient learning of deep networks from decentralised data. *Proceedings of AISTATS*, 54, 1273-1282.
- Vincent, J.L. and Moreno, R. (2010). Clinical review: scoring systems in the critically ill. *Critical Care*, 14(2), 207.
- Horwitz, L.I. et al. (2021). Trends in adverse events during hospitalisation for COVID-19. *Journal of the American Geriatrics Society*, 69(7), 1796-1804.
- Sendak, M. et al. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care. *NEJM Catalyst*, 1(2).
- Ghassemi, M. et al. (2021). False hope: on the limits of health care AI predictions and promises. *PLOS Medicine*, 18(4), e1003466.
- Moons, K.G.M. et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis. *Annals of Internal Medicine*, 162(1), W1-W73.

(SIMI-REB-2025-0018). Local ethics confirmation was obtained at each participating site.

Declarations

Funding

This study was supported by the Swiss National Science Foundation (SNSF grant 200021-224089) and the German Research Foundation (DFG grant SC-4892/3-1). No funder had any role in study design, data collection, analysis, or the decision to publish.

Conflict of Interest

The authors declare no competing interests. No financial relationships exist with clinical AI monitoring software providers, hospital information technology companies, or device manufacturers.

Data Availability Statement

The APSMAF framework and PSES calculator are available at apsmaf-monitor.eu under CC BY 4.0. Aggregate performance data are reported in Tables 3 and 4. Site-level patient data are subject to hospital data governance agreements; requests should be directed to the corresponding author.

Ethical Approval

The study used retrospective clinical monitoring data from existing hospital surveillance systems. Swiss Institute of Machine Intelligence Research Ethics Board approved the study

Appendix A

AI Monitoring Approach Specifications and PSES Scoring Rubric

This appendix provides technical specifications for each AI predictive safety monitoring approach and the complete PSES scoring methodology.