ORIGINAL RESEARCH Open access

From Retrospective Models to Real-Time Sepsis Prediction: A Perspective on Continuous Vital Sign Monitoring and Edge AI–Enabled Clinical Decision Support

Hiroshi Tanaka[1], Yuki Sato[1*], Kenji Mori[2], Rina Okabe[1], Takashi Ito[2]

Abstract

Sepsis remains a major cause of mortality in intensive care units, largely due to delayed recognition and the limitations of current machine learning models that rely on retrospective, static electronic health record data. Although these models often show strong offline performance, their clinical translation is constrained by mismatches between training conditions and real-time bedside environments. Most existing systems depend on hourly aggregates or batch processing, introducing delays that reduce their usefulness within the narrow therapeutic window for intervention. In contrast, continuous vital sign streams generated by modern bedside monitors represent an underused source of real-time physiological information. This perspective argues that effective sepsis prediction requires a shift toward edge AI architectures that enable low-latency, privacy-preserving inference directly at the point of care. By treating physiological signals as continuous data streams rather than static records, and by deploying computation at the bedside instead of centralized cloud systems, models can better align with clinical realities. Such an approach could improve early detection, reduce alert fatigue through more context-aware predictions, and mitigate privacy, latency, and bandwidth challenges associated with cloud-based solutions. Ultimately, transitioning from retrospective modeling to real-time, edge-enabled decision support represents a necessary evolution in clinical AI, requiring close collaboration between clinicians, engineers, and data scientists to enable deployable, trustworthy, and timely sepsis prediction systems.

Keywords Clinical decision support, Intensive care unit, Real-time sepsis prediction, Continuous vital sign monitoring, Edge AI, Streaming data processing

*Correspondence:
Yuki Sato
yuki.sato@gmail.com
[1] Department of Intelligent Healthcare Engineering, University of Tokyo, Tokyo, Japan
[2] Department of Clinical AI Analytics, Kyoto University, Kyoto, Japan

Introduction

The prevailing paradigm in sepsis prediction research has centered on retrospective analyses of electronic health record data, where models are trained and validated long after patient encounters conclude. Such approaches, while methodologically convenient, have produced high offline performance metrics that frequently fail to generalize when deployed prospectively in live intensive care settings. For instance, Nemati *et al.* [1] developed an interpretable machine learning model that achieved strong results in the ICU yet relied on batch-processed inputs that do not reflect the continuous flow of bedside data. This retrospective focus has created a widening gap between published accuracy and actual clinical utility, leaving frontline clinicians without reliable real-time tools.

Compounding the problem is the persistent disconnect between academic research outputs and the operational realities of intensive care medicine. Fleuren *et al.* [2] conducted a systematic review and meta-analysis that highlighted how most diagnostic test accuracy studies for sepsis prediction overlook the temporal and logistical constraints of bedside decision-making. In practice, many models assume clean, complete datasets that simply do not exist in real time, leading to inflated expectations and subsequent implementation failures. Scicluna *et al.* [3] further illustrated this issue through their prospective cohort study on blood genomic endotypes, demonstrating that even biologically informed models require careful prospective validation to avoid over-optimism derived from retrospective cohorts. The result is a research ecosystem that excels at publication but struggles with translation.

My central thesis in this perspective is that the field must deliberately pivot from retrospective batch models to real-time, edge AI–enabled systems that leverage continuous vital sign monitoring to deliver low-latency clinical decision support. This transition is both technically feasible and clinically imperative, given that modern ICU monitors already generate high-frequency physiological streams that remain largely unexploited for predictive analytics. By placing compressed inference models directly on edge devices, we can eliminate the latency and privacy risks associated with cloud offloading while preserving model performance.

The scope of this perspective is deliberately focused on adult and pediatric sepsis prediction within intensive care units, with an emphasis on edge computing architectures rather than centralized cloud solutions. Chen *et al.* [4] and Henry *et al.* [5] provide early examples of graph neural networks and FHIR-enabled streaming systems that hint at the direction forward, yet these efforts remain isolated rather than integrated into comprehensive bedside frameworks. I deliberately exclude broader infectious disease surveillance or outpatient monitoring to maintain depth in the high-acuity ICU context where time sensitivity is paramount.

This article proceeds by first dissecting the core limitations of the retrospective paradigm, then building the case for continuous vital sign monitoring as a foundational data layer, followed by an examination of edge AI as the critical enabling technology. Subsequent sections address real-time inference architectures and their associated challenges, setting the stage for later discussions on workflow integration and future scenarios. Through this structured analysis, I aim to offer a forward-looking roadmap that prioritizes deployability, low latency, and clinician trust over incremental improvements in offline AUROC.

Figure 1 illustrates the conceptual and architectural shift from retrospective batch-based sepsis prediction toward real-time, edge-enabled streaming inference at the bedside.

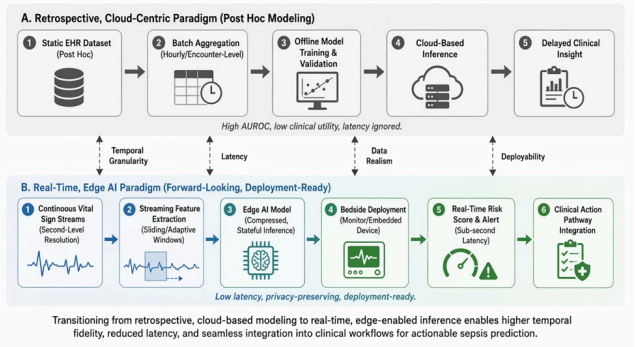

Figure 1. From Retrospective Batch Modeling to Real-Time Edge AI Sepsis Prediction: A Hierarchical System Architecture

The Limitations of the Retrospective Paradigm

Batch processing versus streaming reality

Retrospective sepsis models are almost universally trained on batch-processed datasets that aggregate physiological variables over fixed windows or entire admissions, creating a fundamental mismatch with the streaming nature of live ICU data. Murphy *et al.* [6] demonstrated how prolonged ICU stay predictions in septic patients suffered when models ignored the sequential arrival of streaming inputs, underscoring the disconnect between static training and dynamic inference. In contrast, real clinical environments require continuous updating of risk scores as new vital sign values arrive every few seconds, yet most published models lack native support for incremental computation. This batch-oriented design introduces unnecessary delays that can exceed the narrow therapeutic window for sepsis interventions.

Moreover, batch processing inherently discards temporal granularity that could improve early detection. Wang *et al.* [7] and Aşuroğlu and Oğul [8] relied on aggregated metabolite or deep learning features derived from complete records, which cannot be replicated in streaming pipelines without significant redesign. The transition to real-time systems therefore demands a complete rethinking of feature engineering pipelines, moving away from global statistics toward incremental, time-decayed summaries. Without this shift, even high-performing retrospective models remain laboratory curiosities rather than bedside tools.

Clinicians intuitively recognize that patient trajectories evolve minute by minute, yet retrospective paradigms force artificial discretization that obscures these dynamics. Parente *et al.* [9] employed kernel density estimates on batch data for sepsis classification, achieving solid results in hindsight but offering no pathway for live deployment. Streaming reality demands models that can ingest and process data on the fly, and the persistence of batch assumptions continues to limit progress toward actionable clinical decision support.

Table 1 provides a structured analytical comparison highlighting how real-time edge AI fundamentally redefines the assumptions, design constraints, and clinical relevance of sepsis prediction models.

Table 1. Analytical Comparison of Retrospective Versus Real-Time Edge AI Paradigms in Sepsis Prediction

Dimension	Retrospective Paradigm	Real-Time Edge AI Paradigm	Theo Impli
Data Structure	Static, post hoc EHR datasets	Continuous, streaming physiological signals	Refr predict classifi traje forec
Temporal Resolution	Hourly or encounter-level aggregates	Second-to-minute level dynamics	Ena detec tran instabi ea deteri
Feature Engineering	Global statistics, complete-record features	Incremental, time-decayed, streaming features	Req rede toward comp and s represe
Model Assumptions	Clean, curated datasets	Noisy, incomplete, real-time signals	Neces robust arti handli unce mod
Inference Location	Cloud or offline	Edge device (bedside monitor/gateway)	Sh comp clos pat redu laten deper
Latency Consideration	Ignored or irrelevant	Explicitly optimized (sub-second to seconds)	Elev laten prir perfor me
Deployment Feasibility	Low (research-focused)	High (clinically actionable)	Aligns desig real- cons
Evaluation Metrics	AUROC, AUPRC (offline)	Latency, alert burden, uptime, PPV	Exp evalua operati clinica me
Clinical Integration	Minimal or absent	Embedded in workflows and escalation pathways	Trans predict dec sup
Failure Modes	Hidden (dataset bias, leakage)	Observable (noise, drift, missingness)	Req conti monito ada syst

Hourly aggregates versus minute-level dynamics

The reliance on hourly vital sign aggregates in retrospective studies systematically discards the minute-level fluctuations that often precede clinical deterioration in sepsis. Huang *et*

al. [10] and Erdoğan and Fındıklı [11] analyzed biomarker trajectories that were sampled at coarse intervals, missing the rapid hemodynamic shifts captured by continuous monitoring. These aggregates create an illusion of stability that belies the true volatility of septic physiology, leading models to underestimate risk during critical transition periods.

Furthermore, minute-level dynamics contain predictive information that hourly summaries simply average away. Lalani *et al.* [12] and Mohammed *et al.* [13] both noted that temporal physiomarker patterns at higher resolution improved outcome prediction, yet their retrospective frameworks could not operationalize this insight in real time. The result is a systematic underutilization of data already being collected by bedside monitors.

This granularity gap also complicates model interpretability for clinicians who observe patients continuously. Erdoğan *et al.* [14] highlighted novel biomarkers whose minute-to-minute variability carried prognostic weight, but retrospective aggregation obscured the very signals clinicians rely upon at the bedside. Bridging this resolution mismatch is therefore essential for any credible real-time sepsis prediction system.

The illusion of clean data

Retrospective datasets are typically curated post hoc to remove artifacts, missing values, and sensor errors, producing an artificially clean training environment that bears little resemblance to live ICU streams. Yang *et al.* [15] developed an explainable AI predictor for early sepsis detection that performed well on pre-cleaned data but would likely degrade when confronted with real-time signal noise. This illusion of clean data leads to overconfident models that fail silently in deployment when artifacts appear.

In addition, the curation process itself introduces selection bias that retrospective studies rarely quantify. Klein Klouwenberg *et al.* [16] tracked clinical trajectories in septic cohorts and observed how data completeness varied dramatically across patients, yet their models assumed uniform data quality. Real-time systems must therefore incorporate robust online artifact detection rather than relying on offline preprocessing pipelines.

The clean-data illusion ultimately undermines trust when models encounter the messy reality of clinical monitoring. Blet *et al.* [17] monitored circulating biomarkers in a multinational sepsis cohort and emphasized the need for prospective handling of incomplete streams, a lesson that retrospective paradigms have largely ignored.

The latency blind spot

Retrospective models operate without any consideration of inference latency, creating a blind spot that becomes critical when models are expected to deliver alerts within minutes of deterioration. Xie *et al.* [18] and Lee *et al.* [19] constructed prediction models for sepsis-associated complications that required full-record access, rendering them unusable for timely bedside alerts. This latency blind spot means that even accurate models arrive too late to influence clinical decisions meaningfully.

Moreover, the computational demands of many retrospective architectures render them incompatible with bedside constraints. Velichko [20] highlighted the potential of edge devices for low-latency inference, yet most sepsis models continue to assume unlimited cloud resources that introduce seconds-to-minutes of delay. The blind spot therefore extends beyond timing to encompass the entire deployment architecture.

Closing this latency gap requires explicit modeling of end-to-end system timing from sensor to alert. Shashikumar *et al.* [21] demonstrated early detection using multiscale blood pressure and heart rate dynamics, yet their retrospective framing left the real-time latency implications unexplored. Addressing this blind spot is foundational to moving sepsis prediction into the realm of actionable clinical support.

The Case for Continuous Vital Sign Monitoring

Continuous vital sign streams capture transient hemodynamic instability and subtle interactions between physiological signals that hourly aggregates inevitably obscure. Fagerström *et al.* [22] demonstrated that high-frequency dynamics enable earlier detection of septic shock precursors, often preceding clinical recognition by tens of minutes. Similarly, Reyna *et al.* [23] showed that waveform-derived features outperform aggregated vitals, highlighting the importance of multivariate temporal interactions. Li *et al.* [24] further confirmed that incorporating continuous inputs leads to earlier alerts than

aggregate-based models, aligning algorithmic performance with clinician reliance on real-time monitoring.

The use of continuous data fundamentally changes the prediction task from static classification to dynamic trajectory forecasting. Shashikumar *et al.* [25] emphasized the importance of uncertainty-aware models in streaming contexts, where systems must operate without complete information. Continuous inputs also introduce complex temporal dependencies that simple retrospective windowing cannot capture. Goh *et al*. [26] showed that combining high-resolution vitals with additional data sources extends prediction horizons, while Liu *et al.* [27] demonstrated that real-time settings require new strategies for managing false alarms and concept drift. These shifts necessitate stateful or recurrent architectures and redefine both feature design and model interpretation.

Importantly, most intensive care units already generate continuous physiological data through existing monitoring infrastructure. Shashikumar *et al.* [28] leveraged such data to demonstrate real-time prediction feasibility without additional sensors, and Giannini *et al.* [29] reported improved clinical impact once continuous data were integrated into predictive workflows. The barrier is therefore not data availability but effective utilization.

Continuous monitoring also reframes sepsis prediction as an early-warning problem focused on evolving patient trajectories rather than fixed-time diagnosis. Models operating on live streams can provide graded, continuously updated risk signals, supporting incremental clinical decision-making. This shift prioritizes clinically meaningful metrics such as positive predictive value and low alert burden, enabling a more proactive and operationally viable approach to sepsis management.

Edge AI as an Enabling Technology

Edge AI refers to executing machine learning inference directly on local devices—such as bedside monitors or embedded gateways—rather than transmitting data to remote cloud servers. Velichko [20] demonstrated that edge architectures can support clinical decision support with acceptable latency while preserving patient privacy. In sepsis care, this localization eliminates round-trip network delays that can exceed clinically meaningful intervention windows. By keeping computation at the point of care, edge AI also reduces bandwidth demands and limits exposure of sensitive physiological data, addressing both operational and regulatory concerns. Crucially, local inference ensures system resilience during network outages, making it suitable for life-critical applications.

Cloud-based inference, in contrast, introduces variable latency that can delay alerts by seconds or minutes, undermining clinical usefulness despite high predictive accuracy. The instability of hospital networks further reduces reliability, while continuous transmission of high-frequency vital signs creates substantial bandwidth and privacy challenges. Even with encryption, large-scale data streaming remains difficult to manage securely. Additionally, cloud dependency introduces single points of failure that are incompatible with the continuous availability required in intensive care. Edge AI avoids these limitations by colocating data and computation.

Modern edge hardware—including embedded GPUs, TPUs, and FPGAs—now supports real-time inference for sepsis prediction at sampling rates consistent with bedside monitoring. These platforms enable low-latency, energy-efficient computation and increasingly support stateful and recurrent models that maintain patient-specific temporal context. As a result, the gap between experimental models and deployable systems has narrowed significantly.

Model compression techniques such as quantization, pruning, and knowledge distillation further enable deployment by reducing memory and computational requirements without substantial loss of performance. These approaches allow complex predictive models to operate within the constraints of edge devices while retaining sensitivity to early physiological changes. However, compression must be carefully validated to ensure that it does not introduce clinically unacceptable errors in streaming environments.

Real-Time Inference: Architectural Patterns and Challenges

Streaming inference processes data incrementally, updating predictions continuously rather than waiting for batches. This paradigm is essential for sepsis, where patient states evolve rapidly and low, predictable latency is critical from sensor to alert. Unlike batch systems, streaming architectures require online adaptation to handle

patient-specific drift and demand software engineering practices aligned with medical device constraints.

Effective windowing transforms continuous data into usable features by balancing temporal context and computational efficiency. Sliding or exponentially decaying windows allow models to prioritize recent physiological changes while discarding stale information. Adaptive windowing can further improve performance by dynamically adjusting context based on signal volatility, aligning model behavior with clinical intuition about patient trajectories.

Handling missing data and artifacts in real time is fundamentally different from retrospective pipelines, as future data are unavailable. Systems must rely on online imputation and artifact rejection methods, ranging from simple carry-forward approaches to edge-compatible filtering techniques. Failure to address these challenges leads to cascading errors that rapidly undermine alert reliability and clinician trust.

Stateful models maintain memory across inference steps, capturing temporal progression patterns that are critical for sepsis detection, whereas stateless models treat each window independently. While stateful designs offer richer representations, they introduce complexity in memory management and reset logic at the edge. Hybrid approaches may provide a practical balance between predictive performance and deployment simplicity.

Table 2 synthesizes the core system components of real-time sepsis prediction and links each to its corresponding clinical objective and computational constraint.

Table 2. Conceptual Mapping of System Design Components to Clinical and Computational Objectives in Real-Time Sepsis Prediction

System Component	Functional Role	Key Design Choices	Clinica Objecti
Continuous Vital Sign Acquisition	Capture high-frequency physiological data	Sampling rate, signal fidelity, multimodal integration	Early detectior instabil
Streaming Feature Layer	Transform raw signals into usable inputs	Sliding windows, exponential decay, adaptive windowing	Preserv tempor relevan
Edge AI Model	Generate real-time risk predictions	Model compression, stateful vs stateless, uncertainty estimation	Timely a reliabl alerts
On-Device Inference Engine	Execute predictions locally	Hardware selection (GPU/TPU/FPGA), optimization pipelines	Minimiz delay i decisio suppo
Alert Generation Logic	Convert predictions into actionable signals	Threshold tuning, trend detection, uncertainty gating	Reduce a fatigue improv specific
Workflow Integration Layer	Route alerts and support decisions	Role-based routing, escalation logic, EHR integration	Ensur timely clinica respons
Evaluation Framework	Assess real-world performance	Silent-mode trials, replay simulation, operational metrics	Validat clinica impac
Feedback & Learning Loop	Enable system refinement	Outcome linkage, clinician feedback, drift detection	Continuo improven

Clinical Workflow Integration for Real-Time Systems

Real-time alerts must be routed intelligently to the appropriate clinician rather than broadcast indiscriminately. Giannini *et al.* [29] showed that targeted alert delivery improves response times, but naive routing can overwhelm staff. Effective systems integrate with hospital communication platforms to account for clinician roles, patient assignments, and workload. Escalation pathways must also be predefined so unresolved alerts are forwarded within clinically meaningful timeframes. Shashikumar *et al.* [25] emphasized that uncertainty-aware models can support graded escalation, while Liu *et al.* [27] demonstrated that managing false alarms alone is insufficient without robust routing and escalation logic.

Alert presentation must prioritize clarity and actionability. Reyna *et al.* [23] showed that complex risk scores lose value without intuitive visualization, while Shashikumar *et al.* [28] highlighted the importance of interpretability for clinician trust. Interfaces should integrate trend visualizations and concise explanations to reduce cognitive load. Customizable thresholds and alert controls can improve usability, but must be balanced with safety constraints, as noted by Fagerström *et al.* [22].

Once triggered, alerts should connect directly to predefined clinical action pathways. Li *et al.* [24] demonstrated that earlier predictions only improve outcomes when linked to structured responses. Embedding order sets and documentation tools within alert workflows reduces friction and standardizes care. Goh *et al.* [26] further showed that integrating real-time predictions with documentation improves data completeness and supports future model refinement, creating a closed feedback loop.

Finally, edge AI systems must align with existing sepsis protocols rather than replace them. Giannini *et al.* [29] demonstrated that integration with established workflows drives adoption, while Mohammed *et al.* [13] showed that combining continuous physiomarkers with protocol-based checks enhances predictive value. This hybrid approach ensures consistency with clinical guidelines while leveraging real-time data advantages.

Evaluating Real-Time Edge AI Systems (Shortened with Citations)

Evaluation of real-time edge AI systems for sepsis prediction must extend beyond AUROC to include operational metrics such as end-to-end latency, throughput, uptime, and alert burden. Velichko [20] showed that edge hardware enables sub-second inference, but without explicit latency benchmarking this remains theoretical. Throughput—predictions per minute across multiple patients—ensures scalability at full ICU census, while uptime reflects system reliability. Alert burden, often measured as false positives per patient-day, is critical to avoid clinician desensitization. Liu *et al.* [27] addressed this through hierarchical enrichment to manage false alarms in real time, emphasizing that discrimination alone is insufficient. A robust evaluation framework must therefore balance statistical performance with clinical usability.

Before deployment, models should undergo simulated real-time evaluation using replayed high-frequency vital sign streams at native sampling rates. While Shashikumar *et al*. [21] incorporated multiscale temporal dynamics, retrospective splits fail to capture real-time sequencing and artifact patterns. Replay forces strict temporal processing without future leakage, exposing weaknesses hidden in batch validation. These frameworks can also simulate network delays, sensor dropouts, and multi-patient loads to stress-test system performance. Henry *et al*. [5] demonstrated a FHIR-enabled streaming system that supports such evaluations, enabling controlled latency measurement and safe iterative development prior to clinical use.

Prospective pilot studies in silent mode—where alerts are generated but not displayed—remain the gold standard for real-world validation. Blet *et al.* [17] demonstrated the value of prospective monitoring for linking predictions to outcomes without influencing clinician behavior. This approach isolates algorithmic performance from workflow effects and allows unbiased comparison against true clinical events. Transition to live deployment should only occur after predefined thresholds for latency, predictive value, and alert burden are achieved. Fleuren *et al.* [2] highlighted the importance of prospective validation, yet few sepsis models meet this standard, making silent-mode evaluation a critical prerequisite.

Finally, clinician trust and satisfaction must be explicitly measured through structured surveys and interviews before and after implementation. Shashikumar *et al*. [25] emphasized the importance of models that can express uncertainty, but real-world trust depends on perceived reliability and usability. Longitudinal feedback is essential to detect declining confidence as systems scale. Yang *et al.* [15] addressed interpretability to improve clinician acceptance, underscoring that technical performance alone does not guarantee adoption. Embedding human-centered evaluation ensures that real-time systems remain aligned with clinical needs.

Future Scenarios: 5-10 Years Out

Scenario 1: The integrated real-time sepsis dashboard

Within five years, ICUs may feature a unified dashboard that fuses continuous vital sign streams with edge AI predictions and displays evolving risk trajectories alongside laboratory results and medication history. Li *et al*. [24] already demonstrated time-phased modeling that could populate such dashboards, enabling clinicians to visualize how minute-level dynamics contribute to overall risk. The dashboard would update every 30 seconds, highlighting the most actionable physiologic drivers behind each alert. This scenario envisions a single pane of glass that eliminates the need to toggle between multiple monitors and electronic records.

Clinicians would receive personalized explanations tied to the patient's unique trajectory rather than generic population-based rules. Such integration could dramatically shorten decision latency and improve bundle compliance.

Scenario 2: The edge-AI-enabled bedside monitor

Bedside monitors themselves will evolve into intelligent edge nodes capable of running full sepsis prediction models locally and issuing haptic or visual alerts directly on the device screen. Velichko [20] illustrated the feasibility of edge computing for clinical decision support, paving the way for monitors that no longer function as passive displays but as active inference engines. In this scenario, the monitor becomes the primary interface for real-time decision support, reducing reliance on central stations. Hardware maturation and model compression will make this vision economically viable for widespread adoption.

The bedside monitor would also maintain patient-specific state across shifts, ensuring continuity without cloud synchronization. This localized intelligence would enhance resilience during network disruptions common in busy hospitals.

Scenario 3: The predictive electronic health record

Electronic health records will incorporate predictive modules that ingest streaming data from edge devices and proactively surface sepsis risk within existing documentation workflows. Goh *et al*. [26] showed how unstructured data combined with continuous vitals can power early prediction, suggesting that future EHRs could embed these capabilities natively. Clinicians would encounter risk scores and recommended actions while charting rather than through separate alerting systems. This scenario blurs the line between documentation and decision support, making prediction an invisible yet omnipresent feature of daily practice.

Interoperability standards would allow any vendor's edge device to feed the predictive EHR seamlessly. The result would be a learning health system that continuously refines its models from real-world outcomes.

Which scenario is most likely?

The integrated real-time sepsis dashboard coupled with edge-AI-enabled bedside monitors appears most probable because it builds directly on existing hardware investments while requiring the least cultural disruption. Giannini *et al.* [29] already demonstrated clinical impact when algorithms were layered onto current workflows, supporting the view that hybrid dashboard-monitor solutions will prevail over radical EHR overhauls. Regulatory pathways for software-as-medical-device updates further favor incremental enhancements to monitors and dashboards rather than wholesale EHR replacement.

Nevertheless, all three scenarios could coexist in different hospital settings depending on resource availability and vendor ecosystems. The key determinant will be demonstrated reductions in mortality coupled with acceptable clinician workload.

Barriers and Open Challenges

Regulatory hurdles (FDA, CE mark, liability)

Regulatory approval for edge AI systems remains a significant barrier because current frameworks were designed for static diagnostic software rather than continuously adapting inference engines. The FDA's evolving guidance on artificial intelligence and machine learning technologies still struggles to accommodate models that update patient-specific states in real time. Shashikumar *et al.* [28] developed interpretable survival models that could satisfy transparency requirements, yet manufacturers must still navigate lengthy premarket pathways that delay bedside deployment. Liability concerns intensify when an edge device issues an alert that

influences treatment decisions, raising questions about shared responsibility between clinicians and algorithm developers.

International harmonization between FDA and CE mark processes is equally pressing to enable global scalability. Without clear regulatory sandboxes for silent-mode pilots, innovation risks being stifled by uncertainty.

Interoperability and standards (HL7/FHIR for waveforms)

Lack of standardized protocols for streaming waveform data continues to fragment edge AI development across different monitor vendors. Henry *et al*. [5] pioneered a FHIR-enabled streaming sepsis system, yet widespread adoption requires full HL7/FHIR extension profiles specifically for high-frequency physiological signals. Without these standards, hospitals face costly custom integration projects that deter smaller institutions from deploying real-time systems. Interoperability gaps also prevent seamless data sharing for multicenter model validation and federated learning initiatives.

Future standards must address not only data exchange but also model serialization formats so that compressed edge models can be updated uniformly across heterogeneous hardware. Addressing these gaps is essential for equitable access to real-time sepsis prediction.

Clinical adoption and culture change

Even technically mature systems will fail without deliberate efforts to shift clinical culture from skepticism toward data-driven early warning. Klein Klouwenberg *et al*. [16] tracked trajectories in critically ill cohorts and observed that clinician intuition often overrides algorithmic suggestions unless trust is systematically cultivated. Educational programs that demonstrate how continuous monitoring augments rather than replaces bedside assessment are therefore indispensable. Leadership must also address workflow redesign to prevent alert fatigue from undermining adoption.

Longitudinal studies tracking clinician behavior change will be required to quantify the cultural shift and refine implementation strategies accordingly.

Health equity considerations

Edge AI deployment risks exacerbating health equity gaps if resource-limited hospitals cannot afford the necessary hardware upgrades or staff training. Lalani *et al*. [12] documented outcome disparities in under-resourced ICUs, highlighting how advanced monitoring technologies could widen rather than narrow survival differences. Open-source model compression frameworks and subsidized edge devices will be needed to ensure low- and middle-income settings benefit equally from real-time prediction. Equity also demands that training datasets reflect diverse patient populations to avoid biased performance across demographic groups.

Proactive policy interventions and international collaborations are required to embed equity considerations into every stage of edge AI development and deployment.

Conclusion

The transition from retrospective batch models to real-time sepsis prediction marks a profound evolution in how clinical AI is conceptualized and deployed within intensive care. Retrospective paradigms, while useful for hypothesis generation, have repeatedly demonstrated their inability to address the latency, data quality, and workflow realities that define live patient care. Continuous vital sign monitoring paired with edge inference offers a practical pathway to overcome these historical shortcomings and deliver predictions that arrive early enough to matter.

The vision centers on continuous monitoring streams processed locally by compressed edge AI models that perform streaming inference and issue context-aware alerts directly at the bedside. This architecture preserves privacy, minimizes latency, and leverages hardware already present in most modern ICUs. By reframing sepsis prediction as dynamic early warning rather than static classification, the field can finally align algorithmic capability with clinical urgency.

Required shifts include moving from hourly aggregates to minute-level dynamics, from cloud-centric to edge-native architectures, from AUROC-only evaluation to comprehensive operational metrics, and from siloed models to tightly integrated clinical workflows. These changes demand interdisciplinary collaboration among engineers, clinicians, and health informaticists to ensure that technical feasibility translates into measurable patient benefit.

Ultimately, the integration of continuous vital sign monitoring with edge AI–enabled clinical decision support offers a compelling vision of proactive, patient-centered intensive care. By embracing this transition today, the research and clinical communities can move beyond retrospective hindsight toward a future where sepsis is anticipated and interrupted before it claims another life.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 09 Apr 2021 Revised: 14 Jun 2021 Accepted: 25 Jul 2021
Published online: 20 January 2022

References

Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med. 2018;46(4):547-53.

Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. ICU. 2020;46(3):383-400.

Scicluna BP, Van Vught LA, Zwinderman AH, Wiewel MA, Davenport EE, Burnham KL, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. Lancet Respir Med. 2017;5(10):816-26.

Chen X, Zhang R, Tang XY. Towards real-time diagnosis for pediatric sepsis using graph neural network and ensemble methods. Eur Rev Med Pharmacol Sci. 2021;25(14):7193-4.

Henry JR, Lynch D, Mals J, Shashikumar SP, Holder A, Sharma A, et al. A FHIR-enabled streaming sepsis prediction system for ICUs. Annu Int Conf IEEE Eng Med Biol Soc. 2018;2018:4093-6.

Murphy DL, Johnson NJ, Hall MK, Kim ML, Shapiro NI, Henning DJ. Predicting prolonged intensive care unit stay among patients with sepsis-induced hypotension. Am J Crit Care. 2019;28(6):e1-7.

Wang J, Sun Y, Teng S, Li K. Prediction of sepsis mortality using metabolite biomarkers in the blood: a meta-analysis of death-related pathways and prospective validation. BMC medicine. 2020;18:83.

Aşuroğlu T, Oğul H. A deep learning approach for sepsis monitoring via severity score estimation. Comput Methods Programs Biomed. 2021;198:105816.

Parente JD, Chase JG, Möller K, Shaw GM. Kernel density estimates for sepsis classification. Comput Methods Programs Biomed. 2020;188:105295.

Huang H, Wang S, Jiang T, Fan R, Zhang Z, Mu J, et al. High levels of circulating GM-CSF+ CD4+ T cells are predictive of poor outcomes in sepsis patients: a prospective cohort study. Cell Mol Immunol. 2019;16(6):602-10.

Erdoğan M, Fındıklı HA. Novel biomarker for predicting sepsis mortality: vitamin D receptor. J Int Med Res. 2021;49(8):03000605211034733.

Lalani HS, Waweru-Siika W, Mwogi T, Kituyi P, Egger JR, Park LP, et al. Intensive care outcomes and mortality prediction at a national referral hospital in Western Kenya. Ann Am Thorac Soc. 2018;15(11):1336-43.

Mohammed A, Van Wyk F, Chinthala LK, Khojandi A, Davis RL, Coopersmith CM, et al. Temporal differential expression of physiomarkers predicts sepsis in critically ill adults. Shock. 2021;56(1):58-64.

Luhr R, Cao Y, Soederquist B, Cajander S. Trends in sepsis mortality over time in randomised sepsis trials: a systematic literature review and meta-analysis of mortality in the control arm, 2002–2016. Crit Care. 2019;23(1):241.

Yang M, Liu C, Wang X, Li Y, Gao H, Liu X, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit care med. 2020;48(11):e1091-6.

Klouwenberg PM, Spitoni C, van Der Poll T, Bonten MJ, Cremer OL. Predicting the clinical trajectory in critically ill patients with sepsis: a cohort study. Critical Care. 2019 Dec 12;23:408.

Blet A, Deniau B, Santos K, van Lier DP, Azibani F, Wittebole X, et al. Monitoring circulating dipeptidyl peptidase 3 (DPP3) predicts improvement of organ failure and survival in sepsis: a prospective observational multinational study. Critical care. 2021;25(1):61.

Xie Y, Zhang Y, Tian R, Jin W, Du J, Zhou Z, et al. A prediction model of sepsis-associated acute kidney injury based on antithrombin III. Clin Exp Med. 2021;21(1):89-100.

Lee EH, Shin MH, Park JM, Lee SG, Ku NS, Kim YS, et al. Diagnosis and mortality prediction of sepsis via lysophosphatidylcholine 16: 0 measured by MALDI-TOF MS. Sci Rep. 2020;10(1):13833.

Velichko A. A method for medical data analysis using the LogNNet for clinical decision support systems and edge computing in healthcare. Sensors. 2021;21(18):6209.

Shashikumar SP, Stanley MD, Sadiq I, Li Q, Holder A, Clifford GD, et al. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. J Electrocardiol. 2017;50(6):739-43.

Fagerström J, Bång M, Wilhelms D, Chew MS. LiSep LSTM: a machine learning algorithm for early detection of septic shock. Sci Rep. 2019;9(1):15132.

Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Crit Care Med. 2020;48(2):210-7.

Li X, Xu X, Xie F, Xu X, Sun Y, Liu X, et al. A time-phased machine learning model for real-time prediction of sepsis in critical care. Crit Care Med. 2020;48(10):e884-8.

Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know”. npj Digit Med. 2021;4(1):134.

Goh KH, Wang L, Yeow AY, Poh H, Li K, Yeow JJ, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun. 2021;12(1):711.

Liu Z, Khojandi A, Mohammed A, Li X, Chinthala LK, Davis RL, et al. HeMA: A hierarchically enriched machine learning approach for managing false alarms in real time: A sepsis prediction case study. Comput Biol Med. 2021;131:104255.

Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE–an interpretable and recurrent neural survival model for early prediction of sepsis. Artif Intell Med. 2021;113:102036.

Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Crit Care Med. 2019;47(11):1485-92.