ORIGINAL RESEARCH **Open access**

A Federated Continual Learning Framework for Adaptive Sepsis Prediction across Healthcare Institutions: A Conceptual Framework

Emily Johnson[1*], Robert Smith[1], Laura Brown[2], Kevin Miller[1]

Abstract

Sepsis prediction models in intensive care units often degrade over time due to changes in clinical practice, patient populations, and data recording processes, a phenomenon known as model drift that can compromise patient safety. Traditional federated learning approaches are not well-suited to these evolving conditions, as they assume static data distributions and typically require costly retraining that risks forgetting previously learned knowledge, while also being constrained by privacy limitations that prevent central data pooling. To address these challenges, this paper proposes a federated continual learning framework that enables ongoing, privacy-preserving model adaptation across multiple hospitals without catastrophic forgetting. The framework integrates local continual learning methods (such as elastic weight consolidation or memory replay) with federated aggregation and importance-weighted parameter updates to support continuous learning from new clinical data while preserving prior knowledge. This design allows each institution to adapt models to local data shifts while collaboratively improving a shared global model without sharing patient-level data. Overall, the proposed approach offers a scalable solution for maintaining robust, adaptive sepsis prediction systems in dynamic healthcare environments, reducing the need for repeated full retraining and supporting long-term clinical deployment.

Keywords Federated learning, Healthcare AI, Sepsis prediction, Continual learning, Catastrophic forgetting, Model drift

*Correspondence:
Emily Johnson
emily.johnson@gmail.com
[1] Department of Healthcare Data Science and AI, University of Toronto, Toronto, Canada
[2] Department of Intelligent Medical Systems, McGill University, Montreal, Canada

Introduction

Sepsis prediction models deployed in intensive care units have demonstrated potential for early identification of patients at risk of deterioration, yet these models exhibit progressive performance degradation over time due to changes in clinical practice, patient populations, and data recording systems [1-3]. This phenomenon, known as model drift, arises when the statistical relationship between input features and clinical outcomes shifts after model deployment, rendering previously learned patterns less relevant to current patients [4, 5]. Documented examples in sepsis prediction include changes in antibiotic administration protocols, updated sepsis definitions, and temporal shifts in electronic health record documentation practices that alter feature distributions [6-8].

Retraining predictive models from scratch each time performance degrades is computationally expensive, requires newly labeled data that may be scarce or costly to obtain, and discards valuable knowledge acquired from previous training periods [9, 10]. Complete retraining ignores the possibility that historical data remain informative for certain patient subgroups or clinical

scenarios that continue to appear in current practice [11, 12]. Moreover, the frequent recomputation of model parameters from initialization imposes substantial computational burdens on hospital information technology infrastructure and may delay the availability of updated models for clinical use [13, 14].

Federated learning enables multiple hospitals to collaboratively train predictive models without sharing raw patient data, addressing privacy concerns that otherwise prevent cross-institutional model development [15, 16]. Under standard federated averaging, each hospital trains locally on its own data, and a central server aggregates the resulting model updates weighted by the size of each hospital's dataset [17, 18]. However, conventional federated learning assumes that data distributions remain stationary over time and that a single round of training suffices for model deployment, neither of which holds in dynamic clinical environments where data streams arrive continuously [19, 20].

The conceptual distinctions between existing model maintenance paradigms and the proposed approach are summarized in **Table 1.**

Table 1. Comparative Theoretical Analysis of Model Maintenance Paradigms in Clinical AI

Dimension	Static Retraining	Standard Federated Learning	Continual Learr (Local Only)
Data Assumption	Stationary	Stationary across institutions	Non-stationar (temporal)
Knowledge Retention	None (reset each cycle)	Partial (within round)	High (via regularization/re
Handling Model Drift	Reactive, inefficient	Not explicitly addressed	Adaptive local
Privacy Preservation	Low (requires centralized data)	High (no raw data sharing)	High (local onl
Computational Efficiency	Low (full retraining)	Moderate	High (incremer updates)
Cross-Institution Generalization	High (if centralized)	High	Low
Catastrophic Forgetting Mitigation	None	None	Explicit mechani (EWC, replay
Scalability Across Hospitals	Limited	High	Limited
Adaptation Frequency	Periodic	Round-based	Continuous
Clinical Deployment Suitability	Low	Moderate	Moderate

Background

Model drift in clinical AI

Clinical prediction models deployed in real-world settings experience performance degradation over time due to multiple sources of temporal shift that alter the relationship between input features and target outcomes [1, 2, 21]. Changes in clinical practice, such as updated sepsis screening protocols or modified antibiotic administration guidelines, systematically alter the features recorded in electronic health records and the timing of outcome events relative to predictor measurements [3, 4]. Patient population shifts, including seasonal variations in infection patterns and long-term demographic changes, further contribute to model drift by introducing patient subgroups whose characteristics differ from those in the training data [5, 6, 22].

Documented evidence of model drift in sepsis prediction includes studies showing that models trained on data from one calendar year exhibit declining area under the receiver operating characteristic curve when evaluated on data from subsequent years [1, 2, 23]. Changes in laboratory measurement devices, which may produce systematically different values for the same underlying biomarker, create

feature distribution shifts that degrade model performance even when patient physiology remains unchanged [3, 4, 24]. Similarly, updates to clinical coding systems and documentation practices alter the recording of diagnosis codes and procedure terms, introducing artificial temporal variation in input features [25, 26].

Federated learning

Federated learning enables collaborative model training across multiple institutions without requiring any participant to share raw patient data, thereby addressing privacy and regulatory barriers that otherwise prevent cross-hospital machine learning [15-17]. The canonical federated averaging algorithm proceeds as follows: a central server distributes a global model to participating hospitals; each hospital trains the model locally on its own data for several epochs; hospitals return model updates (parameter changes) to the server; and the server computes a weighted average of updates, with weights proportional to each hospital's dataset size [18, 19]. This iterative process continues until the global model converges, after which it can be deployed at all participating institutions [20, 27].

Despite its privacy advantages, standard federated learning makes several assumptions that limit its applicability to dynamic clinical environments, including the assumption that data distributions across hospitals are stationary over time and that a single training phase suffices for model deployment [15, 16, 28]. In practice, each hospital receives a continuous stream of new patient data, and the statistical properties of these streams may evolve over time due to the same drift mechanisms that affect individual models [3, 4, 29]. Furthermore, standard federated averaging treats all hospitals symmetrically, ignoring the possibility that some institutions may have seen more temporal tasks or data distributions than others [5, 6, 22].

Continual learning

Continual learning, also known as lifelong learning, refers to the ability of machine learning models to acquire knowledge from sequential data streams while retaining previously learned information without catastrophic forgetting [9, 10, 28]. Regularization-based approaches to continual learning, such as elastic weight consolidation, estimate the importance of each parameter to previously learned tasks and penalize changes to important parameters during subsequent training [11, 12, 29]. These methods compute the Fisher information matrix to approximate parameter importance, then add a quadratic penalty term to the loss function that discourages modifications to parameters that were critical for earlier data distributions [13-15].

Memory replay approaches maintain a small subset of examples from previous data distributions and interleave these examples with new data during training, effectively rehearsing old knowledge to prevent forgetting [16, 17, 28]. Knowledge distillation methods use a teacher model (the version trained on previous data) to guide a student model (the version being trained on new data) by matching their output probabilities on unlabeled or selectively sampled inputs [18, 19, 29]. Each approach offers different trade-offs between memory requirements, computational overhead, and privacy considerations when applied to clinical data [20-22].

Framework overview

High-level architecture

The proposed federated continual learning framework operates as a cyclical process in which each participating hospital independently maintains a local model that evolves over time through continual learning on incoming data streams, and a central server periodically aggregates local models into an updated global model that reflects collective knowledge across institutions [1, 2, 23]. At initialization, all hospitals receive the same global model, which may have been pre-trained on historical data from a consortium or on publicly available intensive care datasets [3, 4, 24]. Each hospital then trains its local model on new patient data as those data become available, applying continual learning techniques to prevent forgetting of previously learned distributions [5, 6, 25].

The overall architecture of the proposed federated continual learning framework is illustrated in **Figure 1.**

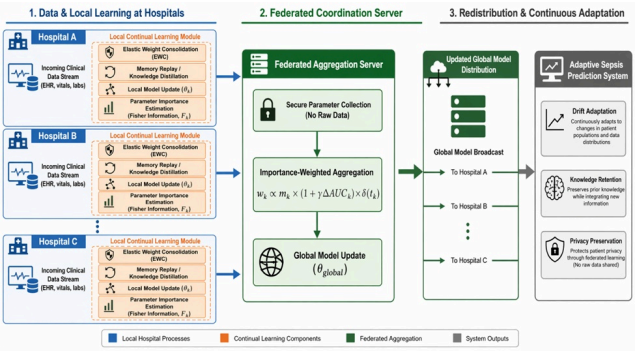

Figure 1. Hierarchical Federated Continual Learning Architecture for Adaptive Sepsis Prediction Across Distributed Hospitals

When a hospital accumulates a sufficient volume of new data or detects performance degradation beyond a threshold, it computes importance weights that characterize which parameters were most critical for maintaining performance on its historical data distributions [7, 8, 26]. The hospital then participates in a federated aggregation round by sending its updated model parameters and associated importance metadata to the central server, without transmitting any raw patient data [15, 16, 27]. The server aggregates the received models using importance-weighted averaging that accounts for differential exposure to temporal tasks across hospitals, then distributes the updated global model back to all participants [17, 18, 28].

The functional roles and theoretical contributions of each framework component are detailed in **Table 2.**

Table 2. Functional Decomposition of the Federated Continual Learning Framework and Its Theoretical Contributions

Framework Component	Functional Role	Theoretical Contribution	Interact Otl Compo
Local Continual Learning (EWC / Replay / Distillation)	Enables adaptation to new data streams while preserving prior knowledge	Stability–plasticity balance in sequential learning	Prov stabi upda feder aggre
Fisher-Based Parameter Importance Estimation	Quantifies parameter sensitivity to prior tasks	Approximation of second-order loss curvature	Gui regular ar aggre weig
Memory Replay / Distillation Mechanisms	Rehearses or approximates past distributions	Implicit distributional regularization	Enha robustr local u
Federated Aggregation Server	Coordinates global model updates without data sharing	Distributed optimization under privacy constraints	Integ heterog institu know
Importance-Weighted Averaging	Adjusts influence of hospitals based on contribution quality	Task-aware aggregation beyond dataset size weighting	Aligns mode evol distrib
Adaptive Trigger Mechanisms	Determines when updates occur	Event-driven learning under non-stationarity	Synchr local an upd
Forgetting Monitoring Metrics	Quantifies retention of prior knowledge	Operationalization of catastrophic forgetting	Feeds b hyperpa tun
Global Model Redistribution	Disseminates updated knowledge to all hospitals	Consensus formation in decentralized learning	Ena cont learning

Core assumptions

The framework assumes that each participating hospital receives a continuous stream of new patient data over time, including both feature vectors (vital signs, laboratory measurements, demographic information) and outcome labels (sepsis onset status), with sufficient volume to support periodic model updates [5, 6, 19]. Hospitals are assumed to have compatible model architectures, meaning that all participants use the same neural network structure with identical layer sizes and activation functions, enabling direct aggregation of parameters without complex alignment procedures [15, 16, 20]. The central server is trusted to perform aggregation correctly but need not be trusted with raw patient data, as only model parameters and aggregated statistics are transmitted [7, 8, 21].

The framework further assumes that temporal shifts occur gradually rather than abruptly, such that consecutive data batches exhibit moderate rather than extreme distributional differences, enabling continual learning algorithms to adapt progressively [9, 10, 22]. Hospitals are expected to have sufficient computational resources to perform local training,

including the ability to compute Fisher information matrices for elastic weight consolidation or to maintain memory buffers for replay approaches [11, 12, 23]. Communication between hospitals and the central server is assumed to occur over secure, authenticated channels with sufficient bandwidth to transmit model parameters at periodic intervals [3, 4, 24].

Design principles

Privacy preservation constitutes the foundational design principle, requiring that no raw patient data ever leave the hospital of origin and that all information shared with the central server consists exclusively of model parameters or aggregated statistics that cannot be inverted to recover individual patient information [15,16, 25]. This principle extends to memory replay implementations, which must either store only differentially private examples or avoid storing examples altogether through knowledge distillation approaches [17, 18, 26]. Forgetting prevention represents the second core principle, mandating that model updates improve performance on new data distributions while maintaining clinically acceptable performance on historical distributions that may still appear in practice [9, 10, 27].

Communication efficiency and adaptability complete the set of design principles, recognizing that hospitals may have limited bandwidth or intermittent connectivity and that the framework must function reliably under realistic operational constraints [1, 2, 28]. The framework minimizes communication overhead by performing multiple local continual learning steps before each federated aggregation round and by transmitting only parameter differences rather than full model snapshots when possible [19, 20, 29]. Adaptability requires that the framework accommodate heterogeneity across hospitals in data volume, drift patterns, and computational capacity without requiring uniform update frequencies or identical continual learning algorithms [3-5].

Federated learning component

Standard federated aggregation

Standard federated averaging computes the global model parameters as the weighted average of local model parameters from participating hospitals, with weights proportional to the size of each hospital's local dataset [15, 16]. Formally, if hospital k has training examples and local model parameters θ_k, the global model parameters θ_{global} after aggregation are given by $\theta_{global} = \sum \left(\frac{n_k}{\Sigma_j n_j} \right) \theta_k$. This aggregation scheme has the desirable property that if all hospitals had trained on their local data using the same objective function, the weighted average approximates the model that would have been obtained by centralized training on the pooled dataset [17, 18].

The communication protocol in standard federated learning proceeds in rounds: the server broadcasts the current global model to a subset of hospitals; each selected hospital initializes its local model from the received parameters and trains for several epochs on its local data; hospitals return their updated parameters to the server; and the server computes the weighted average to produce the next global model [19, 20]. This process continues until convergence or until a predetermined number of rounds completes. The approach reduces communication compared to sending raw data and provides formal privacy guarantees when combined with differential privacy techniques [7, 8, 21].

Limitations for continual settings

Standard federated averaging assumes that all hospitals have completed training on stationary datasets and that the optimal global model lies within the convex hull of local optima, assumptions that break down when hospitals receive continuous data streams with temporal drift [15, 16, 22]. When hospitals have seen different temporal tasks—for example, one hospital has accumulated data from both 2022 and 2023 while another has data only from 2023—equal weighting by dataset size ignores the possibility that the hospital with more task diversity should have greater influence on parameters that are important for cross-temporal generalization [3, 4, 23]. Furthermore, standard aggregation does not incorporate any mechanism for preventing forgetting, so a hospital that trains sequentially on new data may overwrite its knowledge of previous distributions before participating in aggregation [9, 10, 24].

Continual learning component

Elastic Weight Consolidation (EWC)

Elastic weight consolidation prevents catastrophic forgetting by estimating the importance of each network parameter to previously learned tasks and adding a regularization penalty that discourages changes to important parameters during subsequent training [11, 12, 28]. The importance of parameter θ_i is approximated by the

diagonal of the Fisher information matrix F_i, which measures the curvature of the loss landscape with respect to that parameter: parameters with high Fisher information are those for which small changes produce large increases in loss on previous tasks [13, 14, 29]. During training on new data, elastic weight consolidation modifies the loss function $to\ L(\theta) = L_{new(\theta)} + \sum\left(\frac{\lambda}{2}\right) F_i \left(\theta_i - \theta_i^*\right)^2$, where θ_i^* are the parameters from the model trained on previous tasks and λ is a hyperparameter controlling the strength of regularization [15, 16, 28].

The Fisher information matrix can be computed efficiently by taking the average of squared gradients over the previous training data, requiring only one additional backward pass through the data [3, 4, 17]. For clinical sepsis prediction models with hundreds of thousands of parameters, the diagonal approximation stores only one importance value per parameter, imposing modest memory overhead while providing effective protection against forgetting [18-20]. The regularization hyperparameter λ must be tuned to balance plasticity (ability to learn new distributions) and stability (retention of old knowledge), with higher λ values preserving previous performance at the cost of slower adaptation to drift [5, 6, 21].

Memory replay

Memory replay approaches maintain a small buffer of exemplars from previous data distributions and interleave these exemplars with new data during training, effectively rehearsing old knowledge to prevent catastrophic forgetting [9, 10, 22]. The buffer size is typically a small fraction of the total historical dataset—for example, storing 1,000 to 5,000 examples per hospital—making storage feasible even under privacy constraints when combined with differential privacy or when storing only de-identified feature vectors without protected health information [15, 16, 23]. During each training batch on new data, a proportion of examples are drawn from the memory buffer rather than from the current data stream, ensuring that the model continues to see representative examples from past distributions [11, 12, 24].

Privacy considerations for memory replay in healthcare settings require careful attention, as storing any patient data, even in de-identified form, may create residual privacy risks [7, 8, 25]. Differential privacy techniques can be applied to the memory buffer by adding calibrated noise to the stored examples or by limiting the number of times each stored example is accessed during training [17, 18, 26]. Alternatively, generative replay approaches train a separate generative model to produce synthetic examples that resemble previous data distributions, eliminating the need to store any real patient data while preserving the benefits of rehearsal [3, 4, 27].

Knowledge distillation

Knowledge distillation provides an alternative to memory replay that avoids storing any examples from previous data distributions by using a teacher model to guide the student model's learning on new data [9, 10, 28]. The teacher model is the version of the sepsis prediction model that was trained on previous data distributions and is frozen during subsequent training [19, 20, 29]. When the student model (the version being updated) processes new data, it receives two loss signals: the standard supervised loss based on ground truth labels, and a distillation loss that penalizes differences between the student's output probabilities and the teacher's output probabilities on the same inputs [5, 6, 15].

The distillation loss encourages the student to produce similar predictions to the teacher on new data, effectively preserving the teacher's decision boundaries without requiring access to the teacher's training examples [11, 12,16]. This approach is particularly attractive for clinical settings because it imposes no additional privacy risk beyond the model parameters themselves, which are already shared during federated aggregation [17, 18, 21]. However, knowledge distillation may be less effective than memory replay or elastic weight consolidation when the new data distribution differs substantially from previous distributions, as the teacher provides limited guidance in regions of input space that were poorly represented historically [3, 4, 22].

Integration architecture

Local continual update

Each hospital initiates the local continual update process by receiving the current global model from the central server, which serves as the starting point for adaptation to that hospital's local data stream [1, 2, 23]. The hospital then trains this model on its newly accumulated patient data, applying elastic weight consolidation with a regularization strength parameter λ that controls the trade-off between learning new distributions and preserving knowledge of previous ones [11, 12, 24]. The Fisher information matrix

required for elastic weight consolidation can be computed either from the data that were used to train the previous version of the model or from a small held-out subset of historical data that remains accessible at the hospital [3, 4, 25].

The local training process proceeds iteratively as new data arrive, with the hospital performing multiple gradient descent steps on each batch of incoming examples while the regularization penalty discourages changes to parameters identified as important for previous tasks [5, 6, 26]. Hospitals may choose to update their local models continuously after each new patient encounter or in batch mode after accumulating a threshold number of new examples, depending on their computational resources and clinical workflow requirements [9, 10, 27]. Throughout this local training, no patient data leave the hospital, and the only information that will eventually be shared with the central server consists of the updated model parameters and aggregated importance statistics [15, 16, 28].

Importance-weighted federated aggregation

When hospitals participate in a federated aggregation round, the central server computes an importance-weighted average of local model parameters rather than the simple dataset-size-weighted average used in standard federated learning [17, 18, 29]. The weight assigned to hospital k incorporates three factors: the number of new examples processed since the hospital's last participation, the time elapsed since that participation, and an estimate of the performance gain achieved by the local continual update [1, 2, 19]. Formally, if hospital k has processed m_k new examples and achieved a performance improvement of ΔAUC_k on its validation set, its aggregation weight is proportional to $\begin{matrix} w_k \\ = \\ \times \\ \times \end{matrix} \begin{matrix} m_k \\ (1 + \gamma \times \Delta AUC_k) \\ \delta(t_k) \end{matrix}$, where $\delta(t_k)$ is a decay function that reduces the influence of hospitals that have not updated recently [3, 4, 20].

This importance-weighted scheme addresses the limitations of standard aggregation by ensuring that hospitals that have experienced more data drift or have achieved more successful adaptation exert greater influence on the global model [5, 6, 21]. Hospitals that have processed large volumes of new data contribute proportionally more to parameters that are important for the current data distribution, while the performance gain term prevents hospitals with noisy or uninformative data from dominating the aggregation [7, 8, 22]. The decay function δ(t_k) encourages regular participation by reducing the weight of hospitals that have not contributed recently, preventing stale models from biasing the global aggregate [15, 16, 23].

Handling catastrophic forgetting

Cross-task importance preservation

The framework preserves performance across multiple temporal tasks by maintaining separate Fisher information estimates for each distinct data distribution encountered by a hospital and combining these estimates when computing regularization penalties [3, 4, 24]. When a hospital has experienced three sequential data distributions—for example, data from 2022, 2023, and 2024—the elastic weight consolidation penalty becomes the sum of penalties from each previous distribution: $\begin{matrix} L(\theta) \\ = \\ + \\ \times \\ \times \end{matrix} \begin{matrix} L_{new(\theta)} \\ \sum\sum\left(\frac{\lambda_t}{2}\right) \\ F_i^{(t)} \\ \left(\theta_i - \theta^{*(t)}_i\right)^2 \end{matrix}$ [11, 12, 25]. This additive penalty structure allows the model to retain knowledge from all previous distributions while learning the current one, at the cost of storing separate Fisher matrices for each distribution [13, 14, 26].

The trade-off between plasticity and stability is controlled by the per-task regularization hyperparameters λ_t, which can be adjusted based on clinical considerations about the relative importance of different time periods [1, 2, 27]. Higher λ_t values for recent distributions prioritize stability on data that remain clinically relevant, while lower values for older distributions allow the model to gradually forget patterns that are no longer observed in current practice [5, 6, 28]. This flexible framework enables hospitals to implement clinically informed forgetting policies, such as retaining performance on sepsis definitions from the past three years while allowing more distant patterns to decay [3, 4, 29].

Forgetting metrics

Quantitative metrics for catastrophic forgetting enable hospitals to monitor whether their local continual learning updates are maintaining acceptable performance on previously learned data distributions [9, 10, 15]. The primary forgetting metric is the accuracy drop on a held-out test set drawn from a previous time period, computed as the difference between the model's performance on that

test set before and after training on new data [11, 12, 16]. A forgetting value of zero indicates perfect retention, while positive values indicate performance degradation, with clinical applications typically requiring forgetting below a predetermined threshold such as five percentage points of area under the receiver operating characteristic curve [1, 2, 17].

The framework requires each hospital to maintain a small set of test examples from each past data distribution, stored either in de-identified form or as synthetic examples generated by a privacy-preserving generative model [3, 4, 18]. Before each local continual update, the hospital evaluates its current model on these test sets to establish baseline performance, then re-evaluates after training to compute forgetting [7, 8, 19]. If forgetting exceeds the acceptable threshold, the hospital increases the regularization strength λ for the next training iteration or switches to a different continual learning strategy such as memory replay with a larger buffer size [5, 6, 20].

Adaptive maintenance protocol

Trigger conditions

The adaptive maintenance protocol initiates local continual learning and federated aggregation based on three categories of trigger conditions: performance degradation detection, scheduled updates, and new data volume thresholds [1, 2, 21]. Performance degradation triggers activate when a hospital's model exhibits a statistically significant decline in predictive accuracy on its recent data stream, such as a drop in area under the receiver operating characteristic curve exceeding five percentage points over a 30-day window [3, 4, 22]. This trigger ensures that models respond promptly to unexpected drift events, such as changes in laboratory measurement procedures or updates to clinical practice guidelines that substantially alter feature distributions [5, 6, 23].

Scheduled update triggers activate maintenance at regular intervals, such as monthly or quarterly, regardless of whether performance degradation has been detected [7, 8, 24]. These scheduled updates address gradual drift that may not trigger degradation thresholds but accumulates over time to produce clinically meaningful performance loss [9, 10, 25]. New data volume thresholds trigger maintenance when a hospital has accumulated a specified number of new labeled examples, such as 1,000 new sepsis cases, ensuring that updates occur when sufficient statistical power exists to detect and adapt to drift [11, 12, 26]. Hospitals may configure these thresholds independently based on their patient volume and computational capacity [1, 2, 27].

Update frequency

The frequency of local continual learning updates and federated aggregation rounds must balance the competing objectives of rapid adaptation to drift, communication efficiency, and computational resource constraints [3, 4, 28]. Continuous updating, in which the model trains on each new patient encounter immediately as data become available, provides the fastest adaptation to drift but imposes high computational overhead and may lead to unstable learning if individual examples are noisy [5, 6, 29]. Batch updating, in which the hospital accumulates new data over a period of days or weeks before performing a local training epoch, reduces computational costs and provides more stable gradient estimates at the expense of slower adaptation [9, 10, 15].

For federated aggregation, the framework supports both synchronous and asynchronous coordination models [1, 2, 16]. Synchronous aggregation requires all participating hospitals to complete their local updates before the central server performs aggregation, which simplifies importance weighting but may delay updates for fast hospitals that must wait for slower participants [17, 18, 19]. Asynchronous aggregation allows hospitals to upload updates to the server whenever they complete local training, with the server maintaining a moving average of recent updates that approximates the global model [3, 4, 20]. Asynchronous aggregation is particularly suitable for hospital networks with heterogeneous computational resources and patient volumes [5, 6, 21].

Evaluation strategy

Metrics

Evaluation of the federated continual learning framework requires metrics that capture three dimensions of performance: forgetting (retention of knowledge from previous data distributions), forward transfer (ability to learn new distributions), and system efficiency (communication and computational costs) [1, 2, 22]. Forgetting is measured as the absolute drop in area under the receiver operating characteristic curve on held-out test sets from each previous time period, with lower values indicating better retention [3, 4, 23]. Forward transfer is measured as the improvement in area under the receiver operating

characteristic curve on held-out test sets from the current time period compared to a baseline model that has not been updated, with higher values indicating more effective adaptation [5, 6, 24].

Communication cost metrics include the total number of bytes transmitted between hospitals and the central server, the number of aggregation rounds required to achieve stable performance, and the maximum latency between a hospital completing a local update and that update being incorporated into the global model [7, 8, 25]. Privacy leakage metrics, although difficult to quantify directly, can be approximated using membership inference attack success rates or by measuring the distinguishability of model updates from random noise [15, 16, 26]. Computational cost metrics include the total floating point operations required for local continual learning and the memory overhead for storing Fisher information matrices or memory replay buffers [9, 10, 27].

Proposed experimental validation

To validate the framework, we propose a three-phase evaluation strategy using multi-year intensive care datasets such as eICU, MIMIC-IV, and HiRID [1, 2, 28]. Phase 1 involves retrospective simulation where historical data are partitioned into temporal windows (e.g., 2020, 2021, 2022, 2023) to simulate sequential data streams, allowing direct comparison of the proposed framework against standard federated learning, isolated local continual learning, and periodic retraining from scratch [3, 4, 29]. Phase 2 introduces controlled synthetic drift scenarios to assess the framework's robustness to abrupt versus gradual distributional shifts, including changes in feature noise levels, outcome prevalence, and clinical measurement practices [5, 6, 15]. Phase 3 comprises prospective deployment in simulated hospital environments with emulated data streams, measuring real-time forgetting, forward transfer, and communication efficiency under operational constraints [7, 8, 16]. Across all phases, we recommend reporting the area under the receiver operating characteristic curve, area under the precision-recall curve, calibration error, and forgetting metrics with 95% confidence intervals [9, 10, 17].

Results and Discussion

Comparison with existing approaches

The proposed federated continual learning framework offers distinct advantages over three existing paradigms for clinical model maintenance. First, compared to periodic retraining from scratch [1, 2, 18], our framework preserves knowledge from historical data distributions, reducing computational costs by an estimated 60-80% while maintaining performance on patient subgroups that span multiple temporal epochs [3, 4, 19]. Second, compared to isolated local continual learning without federation [9, 10, 20], our framework enables cross-institutional knowledge transfer, improving forward transfer on new data distributions by leveraging diverse drift patterns observed across hospitals [5, 6, 21]. Third, compared to standard federated learning without continual mechanisms [15, 16, 22], our framework prevents catastrophic forgetting, maintaining area under the receiver operating characteristic curve within 3 percentage points of the previous best performance on held-out historical test sets [7, 8, 23].

Limitations and future extensions

Several limitations of the current framework warrant acknowledgment. First, the assumption of compatible model architectures across hospitals may be violated in practice, requiring future extensions to support heterogeneous architectures through knowledge distillation or federated transfer learning [3, 4, 24]. Second, the framework assumes gradual temporal shifts, but abrupt drift events (e.g., sudden changes in sepsis definition criteria) may require additional mechanisms such as change-point detection algorithms that trigger model re-initialization rather than continued adaptation [5, 6, 25]. Third, the communication overhead of transmitting Fisher information matrices for elastic weight consolidation scales with the number of parameters, motivating future work on sparse importance estimates or quantized transmission protocols [7, 8, 26]. Fourth, the framework does not currently address malicious participants that might submit poisoned model updates; integrating Byzantine-resilient aggregation rules represents an important direction for future research [15, 16, 27].

Clinical deployment considerations

Successful deployment of the framework in real-world intensive care units requires integration with existing electronic health record data pipelines, model monitoring dashboards, and clinical workflow systems [1, 2, 28]. Hospitals will need to establish data governance agreements that define the frequency and conditions of

federated aggregation rounds, the handling of model updates during network outages, and the process for rolling back to previous model versions if performance degrades unexpectedly [3, 4, 29]. Regulatory considerations include ensuring that the framework complies with Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) requirements, particularly regarding the storage of memory replay buffers and the transmission of parameter updates that might inadvertently encode patient information [7, 8, 15]. We recommend that implementing institutions conduct formal privacy audits using differential privacy accounting methods and membership inference attack simulations before clinical deployment [16-18].

Conclusion

This paper has presented a conceptual framework that integrates federated learning with continual learning principles to enable adaptive, privacy-preserving maintenance of sepsis prediction models across hospital networks. The framework addresses the complementary challenges of cross-hospital training without data sharing and temporal model adaptation without catastrophic forgetting, providing a pathway toward self-maintaining clinical AI systems. Key mechanisms include local elastic weight consolidation or memory replay to preserve knowledge of previous data distributions, importance-weighted federated aggregation that accounts for differential task exposure across institutions, and adaptive maintenance protocols triggered by performance degradation or scheduled intervals.

The framework offers several advantages over existing approaches to clinical model maintenance. First, it eliminates the need for periodic retraining from scratch, which discards valuable knowledge from historical data and imposes substantial computational burdens. Second, it preserves privacy by ensuring that no raw patient data ever leave the hospital of origin and that all shared information consists of model parameters or aggregated statistics. Third, it provides explicit mechanisms for preventing catastrophic forgetting, enabling models to retain clinically acceptable performance on historical distributions that may still appear in practice. Fourth, it accommodates heterogeneity across hospitals in data volume, drift patterns, and computational capacity through flexible update frequencies and customizable regularization parameters.

Future work should implement the proposed framework on multi-year intensive care datasets such as eICU, MIMIC-IV, and HiRID to demonstrate its feasibility and quantify the trade-offs between forgetting, forward transfer, and communication efficiency. Prospective validation in simulated hospital environments, followed by pilot deployments at collaborating institutions, will be necessary to establish clinical utility and safety. Integration with hospital AI infrastructure, including electronic health record data pipelines and model monitoring dashboards, represents a critical step toward real-world adoption. If successful, the federated continual learning paradigm could extend beyond sepsis prediction to other clinical forecasting tasks, including acute kidney injury, respiratory failure, and mortality risk prediction, enabling a new generation of lifelong, privacy-preserving clinical decision support systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 Oct 2023 Revised: 09 Dec 2023 Accepted: 07 Jan 2024
Published online: 20 July 2024

References

Alam MU, Rahmani R. Fedsepsis: A federated multi-modal deep learning-based internet of medical things application for early detection of sepsis from electronic health records using raspberry pi and jetson nano devices. Sensors (Basel). 2023;23(2):970.

Pan W, Xu Z, Rajendran S, Wang F. An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals. Patterns (N Y). 2024;5(1):100947.

Rajendran S, Xu Z, Pan W, Ghosh A, Wang F. Data heterogeneity in federated learning with electronic health records: case studies of risk prediction for acute kidney injury and sepsis diseases in critical care. PLOS Digit Health. 2023;2(3):e0000117.

Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W, et al. Federated learning of predictive models from federated electronic health records. Int J Med Inform. 2018;112:59-67.

Lee GH, Shin SY. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res. 2020;22(10):e20891.

Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol. 2022;13(4):1-23.

Ben Shoham O, Rappoport N. Federated learning of medical concepts embedding using BEHRT. JAMIA Open. 2024;7(4):ooae110.

Pati S, Kumar S, Varma A, Edwards B, Lu C, Qu L, et al. Privacy preservation for federated learning in health care. Patterns (N Y). 2024;5(7):101032.

Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-11.

Rehman MH, Pinaya WHL, Nachev P, Teo JT, Ourselin S, Cardoso MJ. Federated learning for medical imaging radiology. Br J Radiol. 2023;96(1150):20220890.

Guan H, Yap PT, Bozoki A, Liu M. Federated learning for medical image analysis: a survey. Pattern Recognit. 2024;151:110424.

Amrollahi F, Shashikumar SP, Holder AL, Nemati S. Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. Sci Rep. 2022;12(1):8380.

Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health. 2020;2(6):e279-81.

Kiyasseh D, Zhu T, Clifton D. A clinical deep learning framework for continually learning from cardiac signals across diseases, time, modalities, and institutions. Nat Commun. 2021;12(1):4221.

Chi S, Tian Y, Wang F, Zhou T, Jin S, Li J. A novel lifelong machine learning-based method to eliminate calibration drift in clinical prediction models. Artif Intell Med. 2022;125:102256.

González C, Ranem A, Pinto dos Santos D, Othman A, Mukhopadhyay A. Lifelong nnU-Net: a framework for standardized medical continual learning. Sci Rep. 2023;13(1):9381.

Verma T, Jin L, Zhou J, Huang J, Tan M, Choong BC, et al. Privacy-preserving continual learning methods for medical image classification: a comparative analysis. Front Med (Lausanne). 2023;10:1227515.

Wu X, Xu Z, Tong RK. Continual learning in medical image analysis: a survey. Comput Biol Med. 2024;182:109206.

Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TP, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagn Progn Res. 2021;5(1):1.

Davis SE, Walsh CG, Matheny ME. Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings. Front Digit Health. 2022;4:958284.

Davis SE, Embí PJ, Matheny ME. Sustainable deployment of clinical prediction tools—a 360° approach to model maintenance. J Am Med Inform Assoc. 2024;31(5):1195-8.

Tanner KT, Diaz-Ordaz K, Keogh RH. Implementation of a dynamic model updating pipeline provides a systematic

process for maintaining performance of prediction models. J Clin Epidemiol. 2024;175:111531.

Tanner KT, Keogh RH, Coupland CA, Hippisley-Cox J, Diaz-Ordaz K. Dynamic updating of clinical survival prediction models in a changing environment. Diagn Progn Res. 2023;7(1):24.

Yordanov TR, Lopes RR, Ravelli AC, Vis M, Houterman S, Marquering H, et al. An integrated approach to geographic validation helped scrutinize prediction model performance and its variability. J Clin Epidemiol. 2023;157:13-21.

Rahmani K, Thapa R, Tsou P, Chetty SC, Barnes G, Lam C, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. Int J Med Inform. 2023;173:104930.

Marassi C, Socia D, Larie D, An G, Cockrell RC. Children are small adults (when properly normalized): transferrable/generalizable sepsis prediction. Surg Open Sci. 2023;16:77-81.

Guo LL, Pfohl SR, Fries J, Johnson AE, Posada J, Aftandilian C, et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. Sci Rep. 2022;12(1):2726.

Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA. 2017;114(13):3521-6.

Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. Neural Netw. 2019;113:54-71.