Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Privacy-preserving collaborative learning for ICU admission prediction across heterogeneous hospital networks

Aline Rodrigheri Ioste

aline.ioste@mensa.org.br

Universidade de São Paulo

Murilo Afonso Robiati Bigoto

Universidade de São Paulo

Alexandre Dias Porto Chiavegatto Filho

Universidade de São Paulo

Marcelo Finger

Universidade de São Paulo

Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11482758/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Privacy-preserving collaborative learning for ICU admission prediction across heterogeneous hospital networks

Aline Rodrigheri Ioste[1*], Murilo Afonso Robiati Bigoto[3], Alexandre Chiavegatto Filho[2], Marcelo Finger[1]

[1*]Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, Cidade Universitária, São Paulo, 05508-090, SP, Brazil.
[2]Department of Epidemiology, School of Public Health, University of São Paulo, Avenida Dr. Arnaldo, 715, São Paulo, 01246-904, SP, Brazil.
[3]Bioinformatics Program, University of São Paulo, São Paulo, SP, Brazil.

*Corresponding author(s). E-mail(s): aline.ioste@mensa.org.br;
Contributing authors: murilo.bigoto@usp.br; alexdiasporto@usp.br; mfinger@ime.usp.br;

Abstract

Federated learning (FL) enables hospitals to develop predictive models collaboratively without pooling patient-level data, but its generalization to unseen institutions relative to centralized learning remains uncertain. We compared five FL methods with centralized learning for predicting ICU admission using data from 17,287 patients across 23 Brazilian hospitals. Nineteen hospitals participated in model development, and four were held out for out-of-distribution (OOD) evaluation. The centralized model achieved the highest in-distribution ROC–AUC (0.952), but its OOD ROC–AUC was 0.433. Four of the five FL methods had numerically higher OOD ROC–AUC values than the centralized model, although performance varied across methods. CoLN achieved the highest OOD ROC–AUC (0.767), followed by FedAvg (0.698), FedProx (0.629), and FedDyn (0.465); SCAFFOLD achieved 0.356. CoLN also showed the smallest reduction from in-distribution to OOD performance (0.147), and its OOD ROC–AUC was higher than that of each other evaluated method in pairwise DeLong tests with Holm–Bonferroni correction (adjusted $\boldsymbol{p < 0.001}$). These findings show that strong performance in hospitals represented during development may not translate to unseen institutions, and that OOD performance varies across FL methods.

In this cohort, most evaluated FL methods retained higher discrimination in unseen hospitals than centralized learning.

Keywords: Federated learning; non-convex aggregation; collaborative learning; healthcare prediction; intensive care unit admission; multicentre hospitals; electronic health records

1 Introduction

Clinical prediction models based on electronic health records are increasingly used to support risk stratification, early detection of clinical deterioration, and resource allocation in hospital settings [1–3]. In intensive care, predicting admission to the intensive care unit (ICU) is clinically relevant because delayed identification of high-risk patients may compromise outcomes, whereas excessive false-positive predictions may contribute to unnecessary use of limited resources. However, constructing models that remain reliable across institutions requires accounting for differences in patient populations, clinical practices, and data-generating processes.

The pooling of patient-level data from multiple institutions can increase the sample size and population diversity, but the development of centralized models is often difficult in healthcare because clinical records are subject to privacy, ethical, legal, and governance constraints [4, 5]. Clinical data are also frequently fragmented across institutions, stored in heterogeneous formats, and generated under different documentation practices and regulatory environments [6–8]. These constraints limit data sharing and complicate conventional multicentre model development.

Federated learning (FL) provides an alternative in which institutions train models collaboratively without exchanging raw patient-level data. Each institution performs local optimization and shares model parameters or updates that are combined to build a global model. Federated Averaging (FedAvg), for example, performs weighted averaging of local model parameters and provides a simple and communication-efficient framework for distributed learning [9]. However, clinical data are rarely independent and identically distributed across institutions, which remains a central challenge. Differences in sample size, outcome prevalence, patient case mix, laboratory availability, treatment protocols, and data recording practices can produce substantial variation in local model updates and affect the transportability of the resulting global model.

Several federated optimization methods have been developed to address statistical heterogeneity. FedProx introduces proximal regularization to constrain local updates [10]; SCAFFOLD uses control variates to reduce client drift [11]; and FedDyn introduces dynamic regularization to address heterogeneity across clients [12]. Nevertheless, generalization beyond the institutions represented during model development remains an important challenge for federated clinical prediction.

Recent evidence from federated learning for COVID-19 mortality prediction across hospitals illustrates both the potential and the limitations of collaborative model development. Federated training improved predictive performance without centralizing patient-level data, with particularly relevant gains for institutions with smaller

cohorts [13]. At the same time, performance varied according to institutional heterogeneity and learning architecture, suggesting that collaborative training alone does not eliminate the effects of differences across participating hospitals. This motivates investigation of aggregation strategies that explicitly account for variation among locally trained models.

Several established federated approaches address heterogeneity primarily through modifications to local optimization or update dynamics. An alternative direction is to modify the aggregation rule itself by incorporating information about the relationships among local models.

The Combined Learning Neural Network (CoLN) was originally proposed as an alternative to classical parameter averaging [14]. Rather than constructing the global model solely through a weighted combination of local parameters, CoLN incorporates relationships among corresponding parameters of locally trained neural networks into the aggregation process. CoLN has subsequently been discussed in the privacy-preserving deep learning literature as a collaborative weight-sharing approach related to both federated learning and weight-based privacy-preserving strategies [15].

In the present study, we evaluate CoLN for ICU admission prediction using electronic health record data from 23 Brazilian hospitals. CoLN is compared with hospital-specific models, centralized learning, FedAvg, FedProx, SCAFFOLD, and FedDyn. Performance is evaluated both in hospitals represented during model development and in an external set composed exclusively of hospitals not used during federated training. The study examines whether the CoLN aggregation strategy preserves predictive discrimination under inter-hospital distribution shift and whether its performance differs from established federated approaches when evaluated in previously unseen institutions.

2 CoLN Aggregation

Building upon the original CoLN formulation [14], we formalize the aggregation rule evaluated in this study. Let H local neural networks share the same architecture and parameterization. CoLN constructs a global model by combining parameter aggregation with distance-based information derived from the dispersion of corresponding parameters across local models.

For reference, Federated Averaging (FedAvg) [9] constructs a global parameter vector through weighted averaging,

$$\overline{\mathbf{w}} = \sum_{h=1}^{H} \alpha^h \mathbf{w}^h, \qquad \alpha^h \geq 0, \qquad \sum_{h=1}^{H} \alpha^h = 1, \tag{1}$$

where $\mathbf{w}^h \in \mathbb{R}^d$ denotes the parameter vector of the local model trained at client h, and α^h is a non-negative aggregation coefficient. Under these conditions, the resulting global parameter vector lies within the convex hull of the local parameter vectors,

$$\mathcal{C} = \operatorname{conv}\left\{\mathbf{w}^1, \ldots, \mathbf{w}^H\right\}. \tag{2}$$

CoLN modifies this aggregation rule by introducing a distance-based adjustment derived from the dispersion of corresponding parameters across local models. When

this adjustment is non-zero, the resulting parameter is no longer restricted to a convex combination of the corresponding local parameters.

We consider H local neural networks, $M_1, M_2, \ldots, M_H$, all sharing the same architecture with L layers indexed by $\ell \in \{1, \ldots, L\}$. The parameter vector of layer ℓ in model M_h is denoted by

$$\mathbf{w}_\ell^h = \left(w_{1,\ell}^h, \ldots, w_{M_\ell,\ell}^h\right) \in \mathbb{R}^{M_\ell}, \tag{3}$$

where M_ℓ is the number of trainable parameters, including weights and biases, in layer ℓ, and $w_{i,\ell}^h$ denotes the i-th scalar parameter of that layer in model M_h. In the implementation, layer-specific parameter tensors are flattened before the distance calculations.

The objective of the aggregation procedure is to construct a global model

$$\overline{M} = \{\overline{\mathbf{w}}_\ell\}_{\ell=1}^L . \tag{4}$$

The CoLN aggregation rule uses two complementary measures of inter-model dispersion: a layer-level distance, which summarizes variation across the parameters of an entire layer, and a parameter-level distance, which measures variation in individual corresponding parameters across clients.

Layer-level distance.

Let $\mathbf{w}_\ell^h \in \mathbb{R}^{M_\ell}$ denote the parameter vector of layer ℓ in model M_h. The layer-level distance is defined as

$$\mathrm{LDist}(\ell) = \frac{1}{M_\ell} \sqrt{\sum_{i=1}^{M_\ell} \sum_{j<k} \left(w_{i,\ell}^j - w_{i,\ell}^k\right)^2}. \tag{5}$$

The summation over $j < k$ includes all unordered pairs of local models, with $j, k \in \{1, \ldots, H\}$. Thus, $\mathrm{LDist}(\ell)$ summarizes the dispersion of corresponding parameters within layer ℓ across the participating models. The normalization by M_ℓ is applied after summing the pairwise squared differences and therefore defines a layer-level dispersion measure rather than the arithmetic mean of pairwise Euclidean distances. Larger values indicate greater parameter dispersion across local models within the corresponding layer.

Parameter-level distance.

For an individual parameter i in layer ℓ, CoLN defines

$$\mathrm{WDist}(i, \ell) = \sqrt{\sum_{j<k} \left(r_j w_{i,\ell}^j - r_k w_{i,\ell}^k\right)^2}, \tag{6}$$

where $r_h \geq 0$ represents the relative contribution factor associated with local model M_h. This factor may be defined according to the number of training observations available at each client,

$$r_h = \frac{T_h}{\sum_{u=1}^H T_u}, \tag{7}$$

or assigned uniformly,

$$r_h = \frac{1}{H}. \tag{8}$$

In the experiments reported in this study, uniform values of r_h were used for all participating clients. Under this configuration, $\mathrm{WDist}(i, \ell)$ measures the dispersion of a corresponding parameter across local models after applying equal relative contribution factors. No additional normalization by the number of model pairs is applied.

Distance-based parameter adjustment.

CoLN uses the relationship between parameter-level and layer-level dispersion to define an adjustment coefficient $\beta_{i,\ell}$:

$$\beta_{i,\ell} = \begin{cases} \mathrm{WDist}(i, \ell), & \mathrm{WDist}(i, \ell) < \mathrm{LDist}(\ell), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Accordingly, parameters whose dispersion is lower than the corresponding layer-level dispersion receive a non-zero adjustment, whereas parameters with larger relative dispersion are left unchanged. This rule is based on the hypothesis that parameters showing lower relative variation across clients may encode information that is less sensitive to client-specific variation. The formulation itself does not establish that these parameters are necessarily more predictive or more transferable; this hypothesis is evaluated empirically through the in-distribution and out-of-distribution experiments.

Because $\mathrm{WDist}(i, \ell)$ is non-negative, $\beta_{i,\ell}$ is also non-negative. The adjustment therefore depends on the magnitude of parameter dispersion and does not encode the direction of the discrepancy among local parameters. This property is explicit in the current CoLN formulation and distinguishes the adjustment from gradient- or sign-based update rules.

Aggregation coefficients.

The contribution of each local model to the weighted component of the aggregation is represented by coefficients $\{\alpha^h\}_{h=1}^{H}$. Following the original CoLN formulation and its subsequent parameterization [16], the coefficients are normalized using

$$\alpha^h = \frac{\exp\left(\gamma c_\ell r_h\right)}{\sum_{u=1}^{H} \exp\left(\gamma c_\ell r_u\right)}, \qquad c_\ell > 0, \qquad \gamma > 0. \tag{10}$$

Here, r_h is the relative contribution factor associated with client h, c_ℓ is a scaling factor, and γ controls the concentration of the resulting softmax distribution. By construction,

$$\sum_{h=1}^{H} \alpha^h = 1. \tag{11}$$

The scaling factor may be constant across layers ($c_\ell = c$) or adapted according to layer-level dispersion,

$$c_\ell = \frac{c}{1 + \mathrm{LDist}(\ell)}. \tag{12}$$

When the relative contribution factors r_h differ across clients, c_ℓ and γ modulate the concentration of the softmax aggregation coefficients. However, in the experiments reported in this study, r_h was uniform across clients. Consequently,

$$\alpha^h = \frac{1}{H}, \qquad h = 1, \ldots, H, \tag{13}$$

irrespective of c_ℓ and γ. Thus, under the experimental configuration used here, the weighted component of CoLN corresponds to uniform parameter averaging, and the principal departure from convex averaging is introduced by the distance-based adjustment

$\beta_{i,\ell}$.

CoLN aggregation rule.

For parameter i in layer ℓ, the final aggregated value is defined as

$$\overline{w}_{i,\ell} = \sum_{h=1}^{H} \alpha^h w_{i,\ell}^h + \beta_{i,\ell}. \tag{14}$$

The first term corresponds to normalized aggregation of the local parameters. The second term introduces the distance-based adjustment defined in Equation (9). When $\beta_{i,\ell} = 0$, the aggregation of that parameter reduces to a convex combination of the corresponding local parameters. When $\beta_{i,\ell} > 0$, the resulting value is not, in general, constrained to such a convex combination.

Because $\beta_{i,\ell}$ is non-negative, the current formulation introduces a magnitude-based additive displacement. Its direction is therefore determined by the positive coordinate direction rather than by the sign of the local parameter deviations or by an estimated optimization direction. This characteristic is retained from the CoLN formulation evaluated in this study and should be distinguished from gradient-informed or direction-aware aggregation mechanisms.

The complete procedure is summarized in Algorithm 1.

Geometric interpretation.

Each local neural network can be represented as a point in the high-dimensional parameter space defined by its trainable parameters. The quantities $\mathrm{LDist}(\ell)$ and $\mathrm{WDist}(i, \ell)$ summarize aspects of the dispersion of these points at the layer and individual-parameter levels, respectively. Because these quantities enter the aggregation rule through $\beta_{i,\ell}$, the CoLN operator depends not only on the local parameter values but also on pairwise relationships among them.

For conventional convex averaging, the global parameter vector lies within the convex hull of the local parameter vectors. In CoLN, Equation (14) contains an additional term $\beta_{i,\ell}$. Therefore, whenever $\beta_{i,\ell} > 0$, the corresponding aggregated parameter is no longer restricted to a convex combination of its local values. Across multiple parameters, this permits the aggregated model to lie outside the convex region defined by conventional parameter averaging.

Algorithm 1 CoLN aggregation

Require: Local models $\{M_h\}_{h=1}^{H}$ with L layers; parameter vectors $\mathbf{w}_\ell^h \in \mathbb{R}^{M_\ell}$; relative contribution factors $\{r_h\}$; hyperparameters $c > 0$ and $\gamma > 0$
Ensure: Global model $\overline{M} = \{\overline{\mathbf{w}}_\ell\}_{\ell=1}^{L}$

for $\ell = 1$ to L **do**
  Compute LDist(ℓ) using Equation (5)
  if `use_layer_c` is enabled **then**
    $c_\ell \leftarrow c/(1 + \text{LDist}(\ell))$
  else
    $c_\ell \leftarrow c$
  end if
  Compute $\{\alpha^h\}_{h=1}^{H}$ using Equation (10)
  for $i = 1$ to M_ℓ **do**
    Compute WDist(i, ℓ) using Equation (6)
    Compute $\beta_{i,\ell}$ using Equation (9)
    $\overline{w}_{i,\ell} \leftarrow \sum_{h=1}^{H} \alpha^h w_{i,\ell}^h + \beta_{i,\ell}$
  end for
end for
return $\overline{M}$

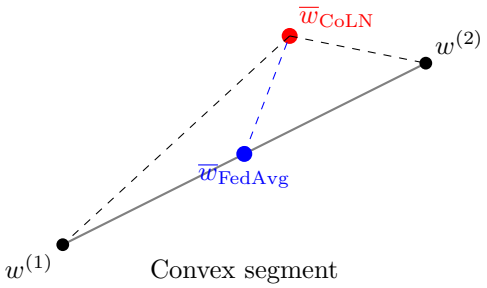

Fig. 1 Illustration of the CoLN aggregation rule. Under convex averaging, the aggregated parameter lies on the convex segment defined by the local parameters. The additive distance-based term used by CoLN can move the aggregated value outside this region. In the current formulation, the adjustment is magnitude-based and does not encode the direction of parameter disagreement.

The geometric interpretation describes the mathematical structure of the aggregation rule and should not be interpreted as establishing that movement outside the convex hull is inherently beneficial. Whether the resulting aggregation improves predictive performance or transportability is an empirical question evaluated in the subsequent experiments.

Relationship to Existing Federated Methods: FedAvg, FedProx, SCAFFOLD, and FedDyn address different aspects of federated optimization. FedAvg combines locally trained models through weighted parameter averaging [9]. FedProx introduces a proximal term into the local objective to constrain local updates

under heterogeneous client distributions [10]. SCAFFOLD uses control variates to reduce client drift during local optimization [11], whereas FedDyn introduces dynamic regularization into the federated optimization procedure [12].

CoLN differs from these approaches in the component modified by the proposed formulation. Rather than introducing an additional local optimization objective or correction term, CoLN modifies the construction of the aggregated parameters by incorporating pairwise parameter dispersion through LDist, WDist, and $\beta_{i,\ell}$. Thus, the aggregation rule depends on both the local parameter values and distance-based quantities derived from the set of participating local models.

Under the uniform-client configuration evaluated in this study, the normalized aggregation coefficients α^h are identical across clients. The experimental formulation can therefore be viewed as uniform parameter averaging followed, for parameters satisfying $\text{WDist}(i, \ell < \text{LDist}(\ell)$, by the additive adjustment $\beta_{i,\ell}$. This distinction is important for interpreting the experiments: the observed differences between CoLN and conventional averaging cannot be attributed to differential client weighting through α^h, but instead arise from the distance-based component of the aggregation rule.

The empirical analyses presented in the following sections evaluate whether this modification is associated with differences in predictive performance under both in-distribution evaluation and inter-hospital distribution shift.

3 Methods

3.1 Study design and data source

We conducted a retrospective multicentre study using electronic health record data from 23 Brazilian hospitals, including public, private, and philanthropic institutions distributed across multiple geographic regions. The objective was to evaluate federated prediction under institutional heterogeneity and, in particular, the transportability of the resulting models to hospitals not represented during model development.

The participating hospitals differed in geographic location, institutional characteristics, sample size, outcome prevalence, and clinical data availability. The geographic distribution of the participating institutions is shown in Figure 2.

3.2 Outcome and predictors

The prediction target was ICU admission during hospitalization. A binary outcome indicated whether a patient was admitted to the ICU. The multicentre dataset comprised 17,287 patient records and included demographic, anthropometric, clinical, physiological, laboratory, COVID-19-related, and institutional variables.

Before model development, candidate variables were screened for eligibility as predictors. Variables containing information occurring after the prediction target, or otherwise considered unavailable at the intended time of prediction, were excluded to reduce the risk of target leakage. Patient and hospitalization identifiers and date-related variables not intended for prediction were also excluded. Raw height and weight

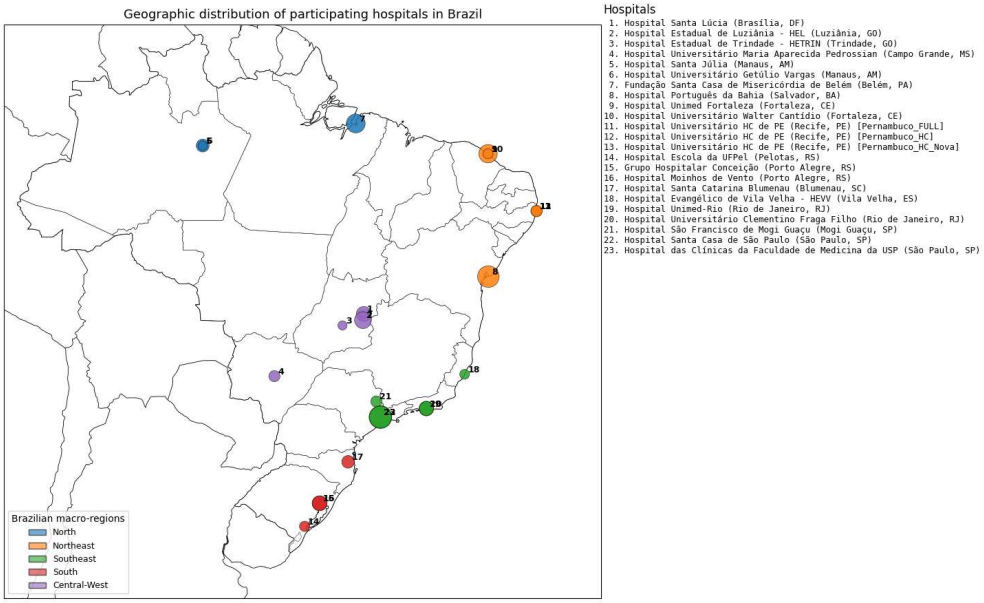

Fig. 2 Geographic distribution of hospitals included in the multicentre cohort. Locations of the 23 participating hospitals across Brazilian macro-regions. Hospitals represented multiple geographic and institutional settings and contributed heterogeneous patient cohorts to the study.

were excluded when redundant with derived body mass index. Only variables retained after this screening procedure were used as inputs to the predictive models.

Table 1 summarizes the candidate variables available in the source dataset before predictor selection. Variables listed in the table therefore represent the pool of candidate variables and should not be interpreted as indicating that every variable was included as an input to the predictive models.

3.3 Ethics approval and consent to participate

All methods were carried out in accordance with relevant guidelines and regulations, including the resolutions of the Brazilian National Health Council (CNS) and the National Research Ethics Commission (CONEP). The study received ethical approval from the University of São Paulo's Institutional Review Board (IRB), under reference number 32872920.4.1001.5421, which included a waiver of informed consent. This approval also covered the use of data and collaboration with all hospitals participating in the IACOV-BR (Artificial Intelligence for COVID-19 in Brazil) network.

Table 1 Candidate variables available in the source multicentre dataset before predictor selection. Variables listed in this table represent the candidate variable pool and were subject to eligibility screening before model development.

Category	Variables available before predictor selection
Demographics and anthropometrics	age, race, body mass index (BMI)
Clinical context	mechanical ventilation (MV), parenteral nutrition, vasoactive drugs, dialysis
Clinical scales	Braden score
Vital signs	heart rate (HR), respiratory rate (RR), systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), temperature, oxygen saturation
Hematology	hemoglobin (Hb), hematocrit (Ht), red blood cells (RBC), platelets, corpuscular indices (MCHC, MCH, RDW, MCV), white blood cell count, neutrophils, lymphocytes, monocytes, basophils, eosinophils, neutrophil-to-lymphocyte ratio (NLR), lymphocyte-to-CRP ratio
Biochemistry and inflammatory markers	C-reactive protein (CRP), albumin, LDH, ALT (TGP), AST (TGO), total bilirubin, direct bilirubin, indirect bilirubin, urea, creatinine, sodium, potassium, magnesium, total calcium, ionized calcium, glucose
Severity and prognosis biomarkers	troponin, D-dimers, BNP, CPK, CKMB, fibrinogen, venous lactate, arterial lactate
Coagulation	INR, aPTT
Arterial blood gases	pH, PaO_2, $PaCO_2$, HCO_3^-, base excess (BE), SaO_2
COVID-19 related	RT-PCR source, RT-PCR result
Institutional	hospital of origin (categorical variable)
Outcome	ICU admission (binary: yes/no)

3.4 Hospital selection and evaluation design

Hospitals were assessed before model development according to sample size and outcome-class representation. Hospitals used as federated clients were required to contain sufficient observations from both outcome classes to permit local model fitting and partitioning of the available observations.

Nineteen of the 23 hospitals met these criteria and were included as federated clients, comprising 11,858 patients. Within these hospitals, patient records were partitioned into development and internal test sets. Twenty percent of the available observations were assigned to the internal test set. Within the remaining development data, 10% were reserved for validation and the remainder were used for model fitting. This resulted in 9,478 observations in the development partition and 2,380 observations in the internal test set.

The four hospitals that did not satisfy the client-level modelling criteria were not used for federated training. Instead, their records were retained as a combined unseen-hospital evaluation set. These hospitals contributed no observations to model fitting, federated aggregation, hyperparameter selection, or any other stage of model development. This evaluation was used to examine model performance under inter-hospital distribution shift.

Accordingly, two complementary evaluation settings were considered: an in-distribution (ID) evaluation using held-out patients from hospitals represented during

model development, and an out-of-distribution (OOD) evaluation using patients from hospitals that were not represented during model development.

3.5 Model development

Neural networks were used as the common base learner for the local, centralized, and federated approaches because CoLN operates on corresponding parameters across models sharing the same architecture. All evaluated approaches therefore used an identical multilayer perceptron architecture. The architecture, optimization procedure, loss function, regularization, batch size, early-stopping criterion, and learning-rate schedule were held constant across the evaluated approaches. This design was used to reduce differences attributable to the underlying predictive model and to focus the comparison on the federated learning procedures.

To address class imbalance during model fitting, minority-class oversampling was applied exclusively to training data, targeting a positive-class proportion of 0.35. For federated approaches, oversampling was performed independently within each client's training partition. Validation, internal test, and unseen-hospital evaluation data retained their observed class distributions.

3.6 Comparison models

CoLN was compared with hospital-specific local models, a centralized model trained on pooled development data from the 19 participating hospitals, and four established federated learning methods: FedAvg [9], FedProx [10], SCAFFOLD [11], and FedDyn [12]. All comparison models used the same base neural network architecture and experimental partitions, with the centralized model serving as a non-federated reference under pooled-data access.

3.7 Federated training

Federated model development was performed using the 19 selected hospitals as clients. Training was conducted for 20 global communication rounds. At the beginning of each round, the current global model was distributed to the participating clients, which performed local training using only their own training observations before returning model parameters or algorithm-specific updates for server-side processing. No patient-level observations were exchanged among hospitals during federated training.

The same experimental configuration was used across federated methods, including client partitions, neural network architecture, communication rounds, local training configuration, class-imbalance handling, preprocessing, and evaluation sets. Algorithm-specific components, including proximal regularization, control variates, dynamic regularization, and the CoLN aggregation rule, were implemented according to the respective methods.

3.8 Model evaluation

Predictive discrimination was evaluated using the area under the receiver operating characteristic curve (ROC–AUC). Performance was evaluated separately in the in-distribution (ID) and out-of-distribution (OOD) settings. The ID evaluation used held-out observations from hospitals represented during model development, whereas the OOD evaluation used the combined unseen-hospital set.

To quantify the change in discrimination under inter-hospital distribution shift, we calculated

$$\Delta\text{ROC-AUC} = \text{ROC-AUC}_{\text{ID}} - \text{ROC-AUC}_{\text{OOD}}. \tag{15}$$

A smaller ΔROC-AUC indicates a smaller reduction in discriminative performance between the ID and OOD settings. We additionally calculated the Robustness Index (RI) as the ratio between OOD and ID ROC–AUC,

$$\text{RI} = \frac{\text{ROC-AUC}_{\text{OOD}}}{\text{ROC-AUC}_{\text{ID}}}. \tag{16}$$

Higher RI values indicate greater retention of discriminative performance in the OOD setting.

Pairwise differences in ROC–AUC on the OOD evaluation set were assessed using DeLong's test. Resulting p-values were adjusted for multiple comparisons using the Holm–Bonferroni procedure, with statistical significance evaluated at a two-sided $\alpha = 0.05$.

4 Results

The ID evaluation included 2,380 patients from the 19 hospitals represented during model development, whereas the OOD evaluation included 5,429 patients from four previously unseen hospitals.

Table 2 summarizes the discriminative performance of the evaluated methods in the in-distribution (ID) and out-of-distribution (OOD) settings.

In the ID evaluation, the centralized model achieved the highest ROC–AUC (0.952), followed by FedProx (0.928), FedAvg (0.918), CoLN (0.914), SCAFFOLD (0.902), and FedDyn (0.836). The difference in ID ROC–AUC between CoLN and FedAvg was 0.004.

In the OOD evaluation, CoLN achieved a ROC–AUC of 0.767, compared with 0.698 for FedAvg, 0.629 for FedProx, 0.465 for FedDyn, 0.433 for the centralized model, and 0.356 for SCAFFOLD. Pairwise DeLong tests comparing CoLN with each of the other evaluated methods on the OOD set yielded Holm–Bonferroni-adjusted p-values below 0.001.

The reduction in ROC–AUC between the ID and OOD evaluations was 0.147 for CoLN, 0.220 for FedAvg, 0.299 for FedProx, 0.371 for FedDyn, 0.519 for the centralized model, and 0.546 for SCAFFOLD. The corresponding Robustness Index (RI) was 0.839 for CoLN, 0.760 for FedAvg, 0.678 for FedProx, 0.556 for FedDyn, 0.455 for the centralized model, and 0.395 for SCAFFOLD.

The unseen-hospital evaluation set comprised 5,429 patients, including 712 positive cases, corresponding to an outcome prevalence of 13.1%. No class rebalancing was applied to the OOD evaluation data.

Table 2 Discriminative performance under inter-hospital distribution shift. ΔROC–AUC represents the absolute reduction from the in-distribution (ID) to the out-of-distribution (OOD) evaluation, while the Robustness Index (RI) is computed as ROC-AUC$_{OOD}$/ROC-AUC$_{ID}$. Higher RI and lower ΔROC–AUC indicate greater retention of discrimination in the OOD setting.

Model	ID	OOD	ΔROC–AUC	RI
CoLN	0.914	0.767	**0.147**	**0.839**
FedAvg	0.918	0.698	0.220	0.760
FedProx	0.928	0.629	0.299	0.678
FedDyn	0.836	0.465	0.371	0.556
Centralized	0.952	0.433	0.519	0.455
SCAFFOLD	0.902	0.356	0.546	0.395

Figure 3 shows the change in ROC–AUC between the ID and OOD evaluations for each method. The magnitude of the reduction varied across approaches, ranging from 0.147 for CoLN to 0.546 for SCAFFOLD.

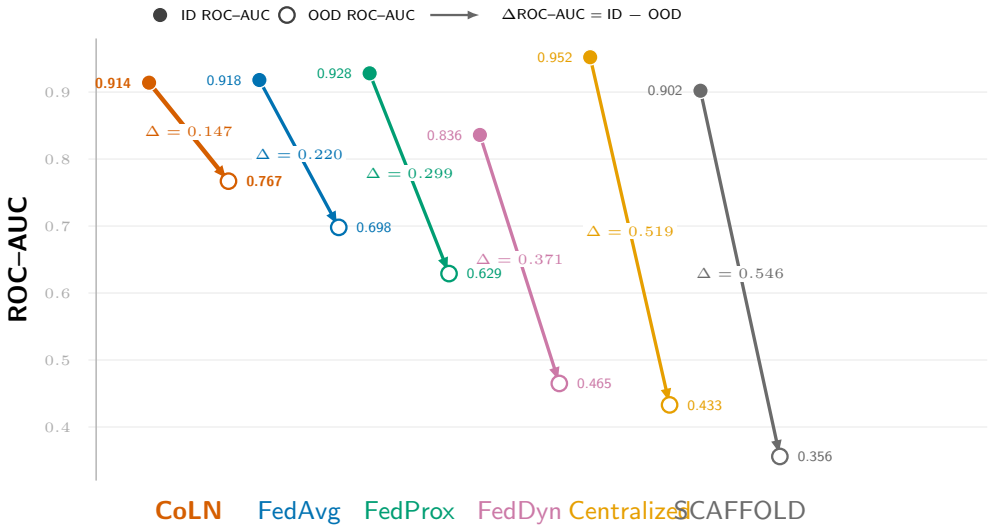

Fig. 3 Discriminative performance under inter-hospital distribution shift. Filled markers indicate ROC–AUC in the in-distribution (ID) evaluation, and open markers indicate ROC–AUC in the out-of-distribution (OOD) evaluation using previously unseen hospitals. Arrows connect the ID and OOD performance of each model, with their length representing the absolute reduction in discrimination, ΔROC-AUC = ROC-AUC$_{ID}$ − ROC-AUC$_{OOD}$.

5 Discussion

In this study, we observed a marked contrast between predictive performance within the development distribution and generalization to hospitals not represented during model development. Although the centralized model achieved the highest discrimination in the in-distribution (ID) evaluation, its performance decreased substantially in the out-of-distribution (OOD) setting, whereas CoLN exhibited the smallest reduction in discrimination. These findings indicate that strong performance within the development distribution does not necessarily translate into cross-institutional generalization.

This pattern is particularly relevant in multicentre clinical data, where differences in patient populations, disease prevalence, laboratory procedures, admission policies, and clinical workflows can produce substantial inter-hospital heterogeneity. Models trained within participating institutions may therefore capture statistical relationships that are not preserved when the institutional distribution changes. In our experiments, this became evident when models with strong ID discrimination showed markedly different behavior in previously unseen hospitals.

The behavior observed with CoLN suggests that the aggregation strategy may contribute to these differences in generalization. Unlike weighted averaging alone, CoLN incorporates layer-level and parameter-level distances among local models during aggregation. We hypothesize that this mechanism may favor parameter configurations associated with greater agreement across participating hospitals and, consequently, with relationships that are more consistently shared across institutions. Although the present experiments do not establish this mechanism causally, the hypothesis is consistent with the smaller degradation observed with CoLN and can be directly examined through ablation analyses.

The experimental design also connects federated learning with Domain Generalization, as four hospitals were intentionally excluded from model development and used as previously unseen clinical domains. CoLN does not explicitly optimize for domain invariance, but its greater retention of discrimination under this setting indicates that the aggregation process may be relevant to model behavior beyond the participating federation. This distinction is clinically relevant because federated training enables collaboration without centralizing raw patient-level data but does not, by itself, establish transportability to other institutions. Our findings therefore support evaluating federated clinical models not only within participating hospitals but also in institutions excluded from model development.

These findings should be interpreted within the scope of the experimental design. The study addressed a single clinical prediction task, the OOD cohort comprised four hospitals analyzed jointly, and a common neural network architecture was used across methods. CoLN was evaluated as a complete aggregation strategy without independent ablation of its components, and the primary analysis focused on discrimination measured by ROC–AUC.

6 Conclusion

This study demonstrates that the aggregation rule is a relevant component of federated clinical prediction when generalization extends beyond the institutions participating in model development. By incorporating relationships among local model parameters, CoLN provides an alternative to aggregation based exclusively on parameter averaging and showed greater preservation of discrimination under the inter-hospital distribution shift examined here. These findings position aggregation as a component to be considered not only for constructing a global federated model, but also for studying its transportability across heterogeneous healthcare institutions.

Declarations

- Funding Funding for this research was provided by the National Council for Scientific and Technological Development (CNPq) and the Department of Science and Technology of the Secretariat of Science, Technology, Innovation and Health Complex of the Ministry of Health of Brazil (MoH), under Grant Number 445020/2023-7.
  This work was also supported by FAPESP grant 2023/00488-5 (SPIRA) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. It was carried out at the Center for Artificial Intelligence (C4AI-USP), supported by FAPESP grant 2019/07665-4 and by IBM Corporation. The sponsors' involvement in the work was limited to providing funding. Marcelo Finger was partly supported by CNPq grant PQ 302963/2022-7.

- Conflict of interest/Competing interests The authors declare that they have no competing interests.
- Ethics approval and consent to participate The study was approved by the University of São Paulo's Institutional Review Board (IRB), under reference number 32872920.4.1001.5421. The IRB waived the requirement for informed consent. This approval covered the use of data and collaboration with all hospitals participating in the IACOV-BR network.
- Consent for publication Not applicable. This manuscript does not contain personally identifiable information.
- Data availability The clinical datasets analyzed in this study are not publicly available because they contain sensitive patient information and are subject to institutional data-sharing agreements and applicable privacy regulations. Access may be granted upon reasonable request to the IACOV-BR group through the corresponding authors, subject to institutional approval and compliance with applicable data protection requirements.
- Materials availability Not applicable.
- Code availability The implementation code of the CoLN federated learning framework, including the training, evaluation, and interpretability pipelines, is publicly available at https://github.com/AlineIoste/CoLN.
- Author contribution Aline Ioste conceptualized and designed the study, developed the methodology, performed the formal analyses, and drafted the manuscript. Murilo Afonso Robiati Bigoto curated and preprocessed the datasets. Alexandre Chiavegatto Filho and Marcelo Finger supervised the study, contributed to methodological

validation, and critically revised the manuscript for intellectual content. All authors reviewed and approved the final version of the manuscript.

References

[1] Shahid, N., Rappon, T., Berta, W.: Applications of artificial neural networks in health care organizational decision-making: A scoping review. PloS one **14**(2), 0212356 (2019)

[2] Sordo, M.: Introduction to neural networks in healthcare. Open clinical: Knowledge management for medical care (2002)

[3] Zion, I., Ozuomba, S., Asuquo, P.: An overview of neural network architectures for healthcare. In: 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), pp. 1–8 (2020). IEEE

[4] Yadav, N., Pandey, S., Gupta, A., Dudani, P., Gupta, S., Rangarajan, K.: Data privacy in healthcare: In the era of artificial intelligence. Indian Dermatology Online Journal **14**(6), 788–792 (2023)

[5] Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., Qadir, J.: Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine **158**, 106848 (2023)

[6] Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S., Barteit, S.: Artificial intelligence for strengthening healthcare systems in low-and middle-income countries: a systematic scoping review. NPJ digital medicine **5**(1), 162 (2022)

[7] Petrie, S., McLeod, S., Ho, K.: From fragmentation to functionality: Enhancing coherence of digital health integration in health systems. In: Healthcare Management Forum, vol. 38, pp. 120–124 (2025). SAGE Publications Sage CA: Los Angeles, CA

[8] Coutinho-Almeida, J., Cruz-Correia, R.J., Rodrigues, P.P.: Evaluating distributed-learning on real-world obstetrics data: comparing distributed, centralized and local models. Scientific Reports **14**(1), 11128 (2024)

[9] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282 (2017)

[10] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)

[11] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T.:

Scaffold: Stochastic controlled averaging for on-device federated learning. In: International Conference on Machine Learning (ICML), pp. 5132–5143 (2020)

[12] Acar, D.A.E., Zhao, Y., Navarro, R., Mattina, M., Whatmough, P., Saligrama, V.: Federated learning based on dynamic regularization. In: International Conference on Learning Representations (ICLR) (2021)

[13] Wichmann, R.M., Robiati Bigoto, M.A., Chiavegatto Filho, A.D.P.: Federated learning for covid-19 mortality prediction in a multicentric sample of 21 hospitals. PLoS Computational Biology **21**(11), 1013695 (2025) https://doi.org/10.1371/journal.pcbi.1013695

[14] Ioste, A.R., Durham, A.M., Finger, M.: Combined learning of neural network weights for privacy in collaborative tasks. arXiv preprint arXiv:2205.00361 (2022)

[15] Alathori, S., Mohammed, N.A.: Encryption and privacy preserving strategies in deep learning-review. Multidisciplinary International Research Journal, 74–85 (2026)

[16] Ioste, A., Finger, M.: Establishing the parameters of a decentralized neural machine learning model. In: Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional, pp. 401–412. SBC, Porto Alegre, RS, Brasil (2022). https://doi.org/10.5753/eniac.2022.227342 . https://sol.sbc.org.br/index.php/eniac/article/view/22799