Scalability and Saturation in Swarm Learning: A KPI-Driven Analysis on Real-World Clinical Data

Matteo Mantovani
Department of Computer Science
University of Verona
Verona, Verona, Italy
matteo.mantovani@univr.it

Simone Scaglia
Department of Computer Science
University of Verona
Verona, Italy
simone.scaglia@univr.it

Carlo Combi
Department of Computer Science
University of Verona
Verona, Italy
carlo.combi@univr.it

Abstract

The adoption of distributed machine learning in healthcare is rapidly expanding to leverage multicentric clinical data while adhering to strict privacy regulations. While Swarm Learning (SL) offers a fully decentralized alternative to a standard centralized Federated Learning setting, the scalability of such peer-to-peer networks remains underexplored. Existing evaluations predominantly rely on synthetic datasets and standard utility metrics, failing to capture the complex distributions of real-world clinical data and struggling to identify saturation regimes where adding participants offers negligible benefits. In this paper, we present a comprehensive scalability analysis of a SL network using real-world intensive care datasets (MIMIC and *eICU*). To systematically evaluate the network's behavior, we introduce three novel Key Performance Indicators (KPIs) designed to quantify the learning outcome at the node, patient, and swarm levels. Our findings provide a robust evaluative framework to inform the optimal design, sizing, and deployment of future decentralized clinical collaborations.

CCS Concepts

• **Applied computing** → *Health informatics*; • **Computer systems organization** → **Peer-to-peer architectures**.

Keywords

distributed learning, performance analysis, federated learning, swarm learning, EHR, KPI, mimic, *eICU*

ACM Reference Format:

Matteo Mantovani, Simone Scaglia, and Carlo Combi. 2026. Scalability and Saturation in Swarm Learning: A KPI-Driven Analysis on Real-World Clinical Data. In *17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '26), June 30–July 03, 2026, Rende (CS), Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3807503.3817123

1 Introduction

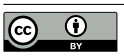
The clinical field has increasingly embraced digital technologies to improve patient care and, more broadly, to transform the organization and experience of healthcare. Large-scale electronic health records, intensive care databases, and imaging repositories are now routinely collected in many hospitals and healthcare networks, and an increasing portion of these resources is made available to researchers under appropriate governance frameworks. Analyzing that data plays a crucial role in decision-making processes, including disease progression forecasting and the design of treatment protocols. However, there are many real-world scenarios (e.g., rare diseases, specialized units, or narrowly defined subpopulations) where the data collected by a single center may not be abundant or diverse enough to support robust model training.

A common solution is to perform learning across multiple centers or servers. However, clinical institutions are often unable or unwilling to share raw data due to privacy regulations, legal constraints, and institutional policies. Privacy-preserving distributed paradigms have been proposed to mitigate these issues, ranging from cryptographic techniques such as homomorphic encryption [7] and Secure Multiparty Computation (SMPC) [26] to Federated Learning (FL) [14], which has emerged as a widely adopted solution. In FL, each site (or client) trains a local model on its own data and periodically exchanges model updates rather than raw records, enabling collaborative training while keeping data in place.

Classical FL architectures, however, typically rely on a central server that orchestrates communication and aggregates local updates. This star-shaped topology introduces a single point of failure and a single point of control, which may be undesirable in cross-institutional healthcare collaborations. To overcome these limitations, Swarm Learning (SL) was proposed in 2021 by Warnat-Herresthal et al. [25]. It removes the permanent central server by introducing a peer-to-peer coordination layer based on a permissioned blockchain and by electing, at each synchronization step, an epoch leader responsible for parameter aggregation among the clients (nodes). Figure 1 illustrates the different learning architectures, highlighting the increased focus on data privacy in federated and swarm learning, as well as the decentralized nature of SL.

A limited number of studies have investigated how these distributed learning systems scale with the number of participating clients; however, many evaluations rely on synthetic benchmarks or modest experimental scales that may not fully reflect large, real-world collaborative deployments [16, 21].

Evaluating these systems using real-world data is critical, as synthetic environments may not capture the complex noise patterns, inherent class imbalances, and unpredictable domain shifts naturally occurring in the clinical environment. This is particularly relevant when assessing the robustness of decentralized learning across both Independent and Identically Distributed (IID) and highly heterogeneous (non-IID) data partitions, which represent the standard operating conditions in real-world collaborative networks [12, 27]. Furthermore, while other studies have addressed this topic,



BCB '26, Rende (CS), Italy

ACM ISBN 979-8-4007-2653-8/26/06
https://doi.org/10.1145/3807503.3817123

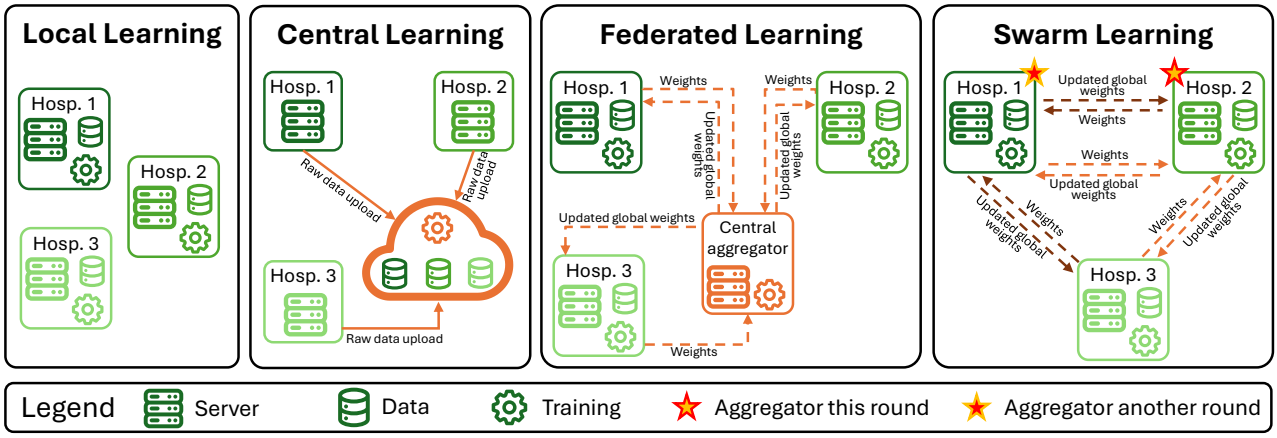


Figure 1: Various learning architectures

they typically report only standard utility and high-level system measures [2, 17].

In [20], the authors provide a custom Key Performance Indicator (KPI) to analyze the behavior of a swarm learning network based on different data distributions, but their study was limited to a few nodes, and it did not focus on the scalability of the network.

Despite these advances, existing evaluations rarely provide dedicated, interpretable KPIs to quantify the impact of adding nodes or patient data to the network. Furthermore, they struggle to identify saturation regimes as the number of nodes or patients grows.

In this paper, we systematically analyze the scalability of a swarm learning architecture to quantify prediction performance as the number of nodes increases, identifying potential plateaus where adding new nodes or patients does not result in significant improvements. Additionally, we investigate whether the granularity of the swarm matters, that is, whether it makes a difference if the same total amount of data (i.e., patients) is concentrated in a few large nodes or split across many smaller ones.

Building on this setting, we introduce a set of KPIs designed to describe how the network behaves as the number of nodes increases. Intuitively, these indicators capture different performance aspects at the node, patient, and network levels, enabling us to quantify not only "how good" the final model is, but also "how much it is still worth" adding additional nodes or patients to the swarm.

Our main contributions are as follows. First, we define three KPIs specifically developed to quantify the learning outcome at the node, patient, swarm, and data levels within the distributed network. Second, we design a comprehensive experimental setting that systematically varies both the number of nodes and the data partitioning strategy (i.e., the number of patients per node), allowing us to verify performance saturation and granularity effects. Third, we apply our methodology to real-world clinical data provided by the *MIMIC-III*, *MIMIC-IV*, and *eICU* intensive care datasets to observe how the swarm learning architecture behaves under both IID and non-IID conditions, and how it performs in terms of the proposed indicators. Finally, we discuss how these findings can inform the design and sizing of future swarm clinical collaborations, helping practitioners decide how many nodes are worth involving and when additional nodes are unlikely to pay off.

For this analysis, we adopt the NVIDIA NVFlare framework, using a blockchain-free approach that isolates the decentralized training and leader-based model merging mechanisms that are relevant to the predictive analysis considered here. This experimental setup does not fully reproduce the original implementation of Swarm Learning, including its permissioned blockchain layer, and the rationale for this choice is discussed in Section 2.2, where we outline the practical limitations of the HPE implementation for large-scale experimentation, and in Section 4.1, where we describe the adopted NVFlare-based system configuration. Accordingly, the present work is centered on predictive scalability and saturation behavior, rather than on broader system-level scalability costs such as communication overhead and runtime, or on the trust and security properties enabled by the permissioned blockchain.

2 Background and Related Work

Swarm Learning (SL) was introduced as a serverless variant of Federated Learning (FL), aiming to remove the permanent central coordinator by relying on decentralized coordination and dynamic leader election for model merging [25].

That study explored the potential of SL for developing disease classifiers using distributed data, but it did not systematically compare SL-trained models with centrally trained ones, nor did it extensively investigate non-uniform data distribution scenarios across nodes. Other studies have since applied SL in clinical contexts, such as Saldanha et al. [23, 24], who trained deep learning models using SL on multicentric gigapixel histopathology datasets, comparing the performance of SL models against models trained on a centralized dataset, but with small and static network topologies.

Regarding scalability, some studies (related to the more traditional FL) vary the number of participating clients to assess performance and resource implications, often reporting standard metrics and coarse system measures (e.g., Area Under the Curve (AUC)/accuracy, loss, CPU/RAM usage) [2, 5, 15]. However, these evaluations rarely provide interpretable indicators that quantify the benefits of adding nodes or patient data, nor do they identify saturation regimes where additional participants may not result in significant improvements. We can find an exception in [20], where the authors introduced a custom KPI to analyze SL performance

under different data distributions, but their study was limited to a few nodes and did not focus on scalability or granularity effects.

The same limitations apply when moving from Electronic Health Record (EHR)-based clinical prediction. Although previous studies have investigated predictive patterns and feature representations from patient-level ICU records [8, 18, 19], node granularity and data partitioning in distributed settings remain unexplored. Additional studies have implemented distributed learning frameworks using multi-hospital Intensive Care Unit (ICU) outcome prediction and cross-site or cross-dataset heterogeneity analyses [3, 9, 28], but they typically fix the network size to a few arbitrary nodes and focus strictly on standard utility metrics. Nevertheless, they do not focus on scalability or granularity effects through KPI suites explicitly designed to capture performance improvements and saturation regimes. This gap motivates KPI-driven methodologies for sizing swarm-like clinical collaborations and for understanding when additional nodes or patients cease to provide meaningful benefits.

2.1 The clinical data sources

2.1.1 MIMIC-III and MIMIC-IV. Medical Information Mart for Intensive Care III (*MIMIC-III*) [11] and *MIMIC-IV* [10] are large-scale, open-access relational databases comprising de-identified clinical data from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts.

MIMIC-III archives data from more than 60 000 admissions corresponding to 46 000 unique patients treated in critical care units between 2001 and 2012. *MIMIC-IV* expands on *MIMIC-III*, featuring updated records from 2008 to 2019. It includes data from nearly 300 000 patients, over 430 000 admissions, and 73 000 Intensive Care Unit (ICU) stays. These databases integrate a comprehensive set of clinical parameters to facilitate the reconstruction of patient medical records, including automated or manually reported vital signs, demographic data, laboratory results, drug prescriptions, medical procedures, diagnostic codes, and caregiver notes. A primary strength of the MIMIC repositories is their high temporal granularity, which captures precisely time-stamped observations and interventions to enable the dynamic modeling of patient trajectories. From an analytical standpoint, these datasets are derived from a single-center environment. As a result, the data reflects a geographically localized and relatively *homogeneous* population, exhibiting characteristics consistent with Independent and Identically Distributed (IID) data structures. Data access is managed through the PhysioNet credentialing process, to ensure that all research activities adhere to rigorous legal and ethical standards for patient privacy while encouraging open scientific inquiry.

2.1.2 eICU. PhysioNet is also responsible for managing the eICU Collaborative Research Database (eICU-CRD, or *eICU* for short) [22], another large-scale, publicly available ICU dataset made available by Philips Healthcare in partnership with the MIT Laboratory for Computational Physiology.

The database comprises de-identified clinical information associated with over 200 000 patient admissions to ICUs between 2014 and 2015. Unlike the single-centric MIMIC, *eICU* includes data from 208 hospitals across the United States, providing a more diverse geographic and demographic sample. Thus, *eICU* represents a multi-center environment, with significant *heterogeneity* regarding patient demographics, clinical care protocols, and data-logging frequencies. This heterogeneity makes the dataset the primary source of Non-Independent and Identically Distributed (Non-IID) data for this work, mirroring the data fragmentation typical of distributed hospital networks. However, similar to the MIMIC datasets, *eICU* includes granular clinical information necessary to reconstruct a patient's clinical history. Thus, extraction and analysis of the same set of features are possible.

2.2 NVIDIA NVFlare software

NVIDIA developed NVIDIA Federated Learning Application Runtime Environment (NVFlare)[1], an open-source framework for distributed learning designed to facilitate collaborative model training across distributed environments without the exchange of raw data. It was initially designed for standard hub-and-spoke federated learning, but now it also provides support to fully decentralized, swarm-like architectures through Peer-to-Peer (P2P) message passing and client-controlled workflows. In this scenario, instead of transmitting updates to a fixed central aggregator, the nodes dynamically elect an epoch leader from among the participating peers at each synchronization round. This leader receives the model parameters from all other nodes, merges them to generate an updated global model, and broadcasts the newly merged weights back to the swarm network. The chosen leader then cedes its role, and the cycle repeats with a newly elected leader until the final training epoch is reached.

Moreover, it is important to underline that NVFlare does not impose restrictions on the maximum number of participating nodes, in contrast with the HPE Swarm Learning Community Edition, which is capped at 5 nodes. NVFlare supports different deployment scenarios, allowing for a single client per physical server, multiple virtualized instances on a single machine, or complex hybrid configurations across heterogeneous infrastructures.

The scalability of NVFlare, along with its support for custom swarm network topologies, enables us to perform a systematic evaluation of saturation points and granularity effects, also with the adoption of our custom KPIs.

3 Methodology

As mentioned in the introduction, the main goal of this work is to analyze the scalability of a swarm learning network by systematically evaluating the performance of the network as the number of nodes increases. Also, we want to investigate whether the granularity of the swarm matters, that is, whether it makes a difference if the same total amount of data (i.e., patients) is concentrated in a few large nodes or split across many smaller ones. In order to properly address these questions, we need to define a set of KPIs that can capture the performance of the network from different perspectives, that is, at the node, patient, and network levels.

First of all, let us introduce some notation:

- $N \in \{2, \ldots, N_{\max}\}$ represents the number of participating nodes.
- m denotes the number of patients contributed by each node.
- $T_N = N \cdot m$ shows the total number of training patients when using N nodes.

[1] https://github.com/NVIDIA/NVFlare

- $M(N)$ indicates the value of a predictive performance metric for the model trained with N nodes.

In each experimental setting, the number of nodes is varied from $N = 2$ to $N = N_{\max}$, where $N_{\max}$ is the maximum number of nodes that can be supported by the dataset given a fixed patient count (i.e., the total number of patients available for training). However, in different experimental configurations, the number of patients per node m can be varied to investigate granularity effects. For example, if we have a total of 12 500 patients available for training, we can set $m = 1250$ to have a maximum of $N_{\max} = 10$ nodes, or we can set $m = 2500$ to have a maximum of $N_{\max} = 5$ nodes. By comparing the performance across these different configurations, we can assess whether the granularity of the swarm (i.e., few large nodes vs many smaller ones) has an impact on the network's performance and scalability.

3.1 Incremental Node Gain (ING)

The *Incremental Node Gain* (ING) quantifies the incremental change in predictive performance when increasing the number of nodes from $N - 1$ to N. It is introduced to systematically evaluate the network's learning trajectory and identify performance plateaus.

It is defined as:

$$\mathrm{ING}(N) = M(N) - M(N-1) \tag{1}$$

ING is expressed in the scale of the metric (e.g., absolute Area Under the Receiver Operating Characteristic (AUROC) units). A positive value indicates improvement due to the addition of the N-th node, whereas values close to zero indicate negligible performance gain, suggesting potential saturation.

3.2 Incremental Gain per Patient ($\mathrm{IG}_{\mathrm{patient}}$)

While the *Incremental Node Gain* (ING) quantifies the absolute improvement in predictive performance attributable to the addition of a new node, it does not account for the varying amount of data contributed by different participants, which can significantly influence the observed gain across different experimental configurations. For instance, an ING of 0.05 in AUROC may be more meaningful if it results from adding a node with 1000 patients rather than one with 4000 patients. To this extent, we introduce the *Incremental Gain per Patient* ($\mathrm{IG}_{\mathrm{patient}}$), which normalizes the performance gain by the number of patients, denoted as m, contributed by the N-th node.

It is defined as:

$$\mathrm{IG}_{\mathrm{patient}}(N) = \frac{\mathrm{ING}(N)}{m}. \tag{2}$$

This indicator reflects the average predictive gain contributed by a single patient from the newly added node, enabling a fair comparison across experiments with varying node sizes and data partitioning strategies. Furthermore, since the incremental gain of a single patient on typical performance metrics (e.g., AUROC) can be extremely small, we also define the Incremental Gain per 1000 Patients (IG_{1000}) to facilitate the interpretation of this metric:

$$\mathrm{IG}_{1000}(N) = \frac{\mathrm{ING}(N)}{m} \times 1000. \tag{3}$$

A progressive decrease in IG_{1000} as the network expands suggests the presence of a data saturation regime, indicating that the informational value of additional patients is diminishing. Conversely, a sustained high value of IG_{1000} across increasing N indicates that each new node continues to provide substantial predictive benefits.

3.3 Node Efficiency Ratio (NER)

To distinguish between intrinsic data saturation and potential efficiency loss due to distributed training, we compare the swarm performance against a central reference.

The baseline is set at $N = 2$, as it represents the minimal node configuration to create a collaborative swarm network.

Let $M_{\mathrm{SL}}(N)$ denote the swarm model performance trained with N nodes, and let $M_{\mathrm{cent}}(N)$ represent the performance of a central model trained on the union of the same T_N patients.

Accordingly, let:

$$\Delta_{\mathrm{SL}}(N) = M_{\mathrm{SL}}(N) - M_{\mathrm{SL}}(2)$$

And

$$\Delta_{\mathrm{cent}}(N) = M_{\mathrm{cent}}(N) - M_{\mathrm{cent}}(2)$$

denote the incremental gain of the swarm and central models, respectively, with respect to the two-node configuration.

We define the *Node Efficiency Ratio* (NER) as the ratio between the incremental gain achieved by the swarm and the corresponding incremental gain of the central model.

However, in cases where the central model reaches saturation, $\Delta_{\mathrm{cent}}(N)$ may approach zero, making the ratio numerically unstable and difficult to interpret. To handle these cases, we use an ε-stabilized denominator. Specifically, when $|\Delta_{\mathrm{cent}}(N)| < \varepsilon$, we replace $\Delta_{\mathrm{cent}}(N)$ with ε or $-\varepsilon$ according to its sign.

Formally, we define the *Node Efficiency Ratio* (NER) as:

$$\mathrm{NER}(N) = \frac{\Delta_{\mathrm{SL}}(N)}{\tilde{\Delta}_{\mathrm{cent}}(N)} \tag{4}$$

where

$$\tilde{\Delta}_{\mathrm{cent}}(N) = \begin{cases} \Delta_{\mathrm{cent}}(N), & \text{if } |\Delta_{\mathrm{cent}}(N)| \geq \varepsilon, \\ \varepsilon, & \text{if } 0 \leq \Delta_{\mathrm{cent}}(N) < \varepsilon, \\ -\varepsilon, & \text{if } -\varepsilon < \Delta_{\mathrm{cent}}(N) < 0 \end{cases}$$

This stabilization avoids poorly interpretable values when the central gain is close to zero, while maintaining the interpretability of the ratio and its sign.

The values of $\mathrm{NER}(N)$ can be interpreted as follows:

- $\mathrm{NER}(N) > 1$: The swarm obtains a higher predictive gain than the central model. This is a rare scenario that can occur, for example, if the SL aggregation acts as a beneficial regularizer.
- $\mathrm{NER}(N) \approx 1$: The swarm scales as efficiently as the central model.
- $0 < \mathrm{NER}(N) < 1$: The swarm improves, but the network is not fully exploiting the additional data compared to the central model.
- $\mathrm{NER}(N) = 0$: The swarm shows no incremental gain relative to the two-node baseline.
- $\mathrm{NER}(N) < 0$: The additional data affect swarm and central learning in opposite ways, indicating that it no longer produces a coherent benefit across the two settings.

3.4 Interpretation of the proposed KPIs

The proposed KPIs offer a comprehensive framework for evaluating the scalability of SL networks across multiple levels.

At the *node level*, ING monitors the incremental change in performance as network size increases.

At the *patient level*, $IG_{patient}$ and IG_{1000} provide a normalized measure of the predictive gain attributable to a fixed number of patients (one or a thousand, respectively). This allows us to evaluate the contribution of the added data independently of the specific node size (m) used in the experiment.

At the *network level*, NER evaluates the overall swarm learning efficiency against a central learning baseline.

Finally, let us note that evaluating these indicators collectively is essential, as the limitations or lack of interpretability of a single KPI under specific experimental conditions are effectively compensated by the complementary information provided by the others.

4 Experimental Evaluation

In this section, we present the experimental evaluation of all the KPIs described in Section 3 using real clinical data from the *MIMIC-III*, *MIMIC-IV*, and *eICU* datasets. Specifically, we examine how the performance of the swarm learning network evolves as we increase the number of nodes, and we investigate whether there are saturation points where adding more nodes does not lead to significant improvements. Moreover, we adopt both IID datasets (i.e., *MIMIC-III* and *MIMIC-IV*) and a non-IID dataset (i.e., *eICU*) to verify the behavior of the swarm learning under different scenarios. A key aspect of our analysis is the impact of data granularity: we systematically investigate whether the swarm's performance is affected when a fixed total number of patients is concentrated into a few large nodes versus being distributed across many smaller ones. There are several experimental configurations that we consider in our evaluation, as shown in Table 1, where the maximum number of nodes ($N_{\max}$) and the corresponding patients per node (m) for each dataset are detailed.

For the IID datasets, we partition the data into training and test sets using a 70-30 split. The training set is then split to create non-overlapping subsets for all the different node configurations. For example, given 12500 patients, we split the data into {5, 10, 20, 40, 80} nodes with {2500, 1250, 625, 312, 156} patients per node, respectively. The test set is unique for each dataset; that is, it is the same for all the nodes involved in the network.

In *eICU*, each node corresponds to a different hospital, and each one of them could contain a different number of patients. We want to reflect the heterogeneity of the original dataset, and to do so, there are three preliminary steps needed. First, for each $N_{\max}$, we have to randomly select only the hospitals that have at least m patients, where m is the number of patients per node in that training configuration. This random selection is performed for each iteration, so the results are not influenced by the specific hospital choices. In the second step, we need to ensure that all nodes have the exact number of patients for each experimental configuration, which is obtained by removing excess patients via stratification. In the third step, we need to create a test set for each node, which is obtained by splitting 30% of the data from each hospital.

Table 1: Summary of all experimental configurations for each dataset, showing the number of training patients ($T_{N_{\max}}$), maximum nodes number ($N_{\max}$), and patients per node (m).

		Number of Nodes ($N_{\max}$)					
Dataset	**Total Patients**	**5**	**10**	**20**	**25**	**40**	**80**
MIMIC-IV	12 500	2500	1250	625	–	312	156
MIMIC-III	12 500	2500	1250	625	–	312	156
eICU	12 500	2500	1250	625	500	–	–

The central learning baseline model is trained and then tested on the union of the training and test sets of all the nodes involved in the swarm network, respectively.

Across all datasets, stratification is used to ensure that the class distribution is preserved across all splits, both at the train-test level and when partitioning the training set into nodes for the IID datasets. In the case of *eICU*, stratification is applied at the hospital level to maintain the original heterogeneity of the data across nodes. The results obtained from the experimental evaluation are averaged over 10 independent training runs, each using a different random seed to shuffle the data.

The learning process is performed using a Feedforward Neural Network (FNN). However, during the incremental expansion of the network from $N = 2$ to $N_{\max}$, the optimal hyperparameters may shift as the swarm aggregates more data. Therefore, within each specific experimental configuration, we perform a grid search every 5 nodes (e.g., at $N = 5, 10, 15$, and so forth). The optimal hyperparameters identified at each step were subsequently applied to the intermediate configurations within that interval; for instance, the parameters optimized for $N = 10$ are also utilized for the intermediate steps $N \in \{6, 7, 8, 9\}$.

The grid search explores learning rates of {0.0001, 0.001, 0.01, 0.1} and batch sizes of {16, 32, 64, 128, 256, 512}, while the number of epochs is fixed to 25. To ensure a fair comparison, an independent grid search over the same hyperparameter space is also performed for both the central and local learning baselines. Moreover, the number of epochs is fixed to 25, and the aggregation round is set to 5 for all the grid search iterations and all the training runs.

Finally, since all nodes have the same size, they all share the same weight in the weighted average aggregation algorithm, equal to the percentage of the training set within each node.

4.1 System configuration and software

We conducted our experiments on a Google Cloud Platform virtual machine with the following specifications: n2-standard-32 instance type, which provides 32 vCPUs, 128 GB of RAM, and 10 GB of SSD storage. We deliberately chose to run the experiments on a machine with a high amount of memory, as the training of the models with a large number of nodes can be computationally intensive. Any systematic evaluation of the computational aspects, such as training time or memory usage, is out of the scope for this paper, and could represent an interesting direction for future work.

To implement the swarm learning framework, we use the NVIDIA NVFlare software v2.7, using a blockchain-free approach while maintaining the core principles of SL, such as decentralized coordination and dynamic leader election for model merging. In this paper, we are interested in the prediction performance of the swarm learning network, and not in the security aspects of the system.

Table 2: Vital signs are discretized into "low" and "high" via thresholds; values in between are "normal", while those outside the ranges are deemed likely erroneous as physiologically implausible. SpO_2 is an exception: as a percentage, 96–100 is "normal" and no "high" category is defined.

	Heart Rate (bpm)	**Diastolic BP** (mmHg)	**Systolic BP** (mmHg)	**White Blood Cells** (K/μL)	**Body Temp.** (°C)	**Resp. Rate** (insp/min)	**SpO_2** (%)
Low	$10 \leq val \leq 60$	$10 \leq val \leq 60$	$50 \leq val \leq 90$	$0 \leq val \leq 4.0$	$15 \leq val \leq 36$	$4 \leq val \leq 12$	$40 \leq val \leq 95$
High	$100 \leq val \leq 300$	$90 \leq val \leq 300$	$140 \leq val \leq 400$	$11.0 \leq val \leq 100$	$38.3 \leq val \leq 50$	$20 \leq val \leq 120$	-

Therefore, the blockchain is not strictly necessary for our analysis, and we can rely on the security features provided by NVFlare without the need for a distributed ledger.

The swarm learning model uses a Feedforward Neural Network (FNN) with one input layer, three hidden layers, and one output layer. The three hidden layers are densely connected, each with 16 nodes and utilizing a rectified linear unit (ReLU) activation function [1]. The output layer is also densely connected and contains a single node with a sigmoid activation function.

The model configuration for training involves the use of the Adam optimization algorithm [13] to effectively minimize the binary cross-entropy loss between the true and predicted labels. This model is consistently applied across all the different experimental configurations, which are the swarm learning, the central learning, and the local learning (i.e., training a model on each node independently without any collaboration).

AUROC was chosen as the primary metric for evaluating the performance of the model, as the task is a binary classification.

4.2 Datasets ETL

In this subsection, we describe the process of Extract, Transform, Load (ETL) performed on *MIMIC-III* 1.4, *MIMIC-IV* 2.2, and *eICU* 2.0 to obtain the datasets used for this paper. The three datasets contain Electronic Health Record (EHR) data from patients admitted to the ICU, and their structure is similar enough to allow us to extract the same set of features for all of them, which is essential for a fair comparison in our experimental evaluation.

Briefly, in this paper, we focused on predicting sepsis diagnosis during the hospital stay, which serves as our class of interest (i.e., *true* if the patient has been diagnosed with sepsis, *false* otherwise). Then, we selected several vital signs plus medications that are known to be heavily or weakly related to sepsis treatments [4, 6].

In detail, we started by extracting only adult patients (age between 18 and 89) with at least one ICU stay of at least one day, at least one diagnosis, and at least three drug administrations during the stay. This choice allowed us to exclude short-term admissions and focus on cases more likely to involve significant interventions, critical care requirements, and more comprehensive clinical data. Let us note that all these filters are applied at the admission level, meaning that if a patient is admitted multiple times, we consider only those admissions that meet the specified criteria, and each admission is treated as a separate instance in the dataset. In this setting, each row of the resulting dataset corresponds to a single admission, while each column contains an element of the patient's medical history selected as a feature for the analysis.

The vital signs considered were heart rate, respiratory rate, oxygen saturation, body temperature, white blood cell count, and systolic and diastolic blood pressure (both as arterial and non-invasive, treated as separate vital signs). These parameters are fundamental indicators of the patients' status, especially when they are diagnosed with sepsis. These vital signs were then transformed into categorical features by discretizing their values into "low", "normal", and "high" categories based on the thresholds indicated in the guidelines [4, 6] and described in Table 2. Values exceeding the defined ranges are considered likely to be incorrectly reported, as they fall outside physiologically plausible limits. For all the vital signs described, we created three different features containing the label for the *minimum*, *maximum*, and *average* measurements recorded during the admission. In *MIMIC-III* and *MIMIC-IV*, the vital signs were extracted from the `chartevents` table, while in *eICU* they were extracted from the `vitalperiodic` (i.e., vital signs captured periodically from the bedside monitors), `vitalaperiodic` (i.e., vital signs captured at irregular basis), and `nursecharting` (i.e., vital signs captured by the nurse) tables. However, since the same vital sign can be recorded in multiple tables, we had to apply specific criteria to select the source of each measurement in *eICU*. For the *average* values, we prioritized the `vitalperiodic` table, as it typically contains more samples for most parameters. For the *minimum* and *maximum* values, we prioritized the `nursecharting` table over the `vitalperiodic` and `vitalaperiodic` tables as they are more likely to be validated and to avoid anomalous spikes. Finally, for temperature and White Blood Cell count (WBC), we consistently used the `nursecharting` and `lab` tables, respectively, for all three values (min, mean, max) due to their specific nature as a vital sign and a laboratory test.

We then focused on the drug administration by selecting a set of drugs usually recommended to treat sepsis by the guidelines [4, 6] along with weakly sepsis-related medications. The drugs considered include vasopressors (norepinephrine, epinephrine, phenylephrine, vasopressin, dopamine), broad-spectrum antibiotics (vancomycin, piperacillin/tazobactam), potassium chloride for electrolyte correction, proton pump inhibitors (omeprazole, pantoprazole), and the beta-blocker metoprolol. Ciprofloxacin and acyclovir were also included due to their occasional, non-standard use. While some of these medications are central to sepsis management, others (e.g., metoprolol, potassium chloride, phenylephrine, omeprazole, pantoprazole) have broader clinical indications. Each drug is treated as a binary feature, indicating whether it was administered at any point during the hospital stay.

Lastly, we used the ICD-9 codes to identify sepsis and severe sepsis cases as our target variable. At the end, all three resulting datasets contain the same 41 features, ensuring a fair comparison across different experimental configurations and datasets. *MIMIC-III* dataset is composed of 17 569 admissions, with a sepsis prevalence of 13.7%, while *MIMIC-IV* contains 51 100 admissions with a sepsis

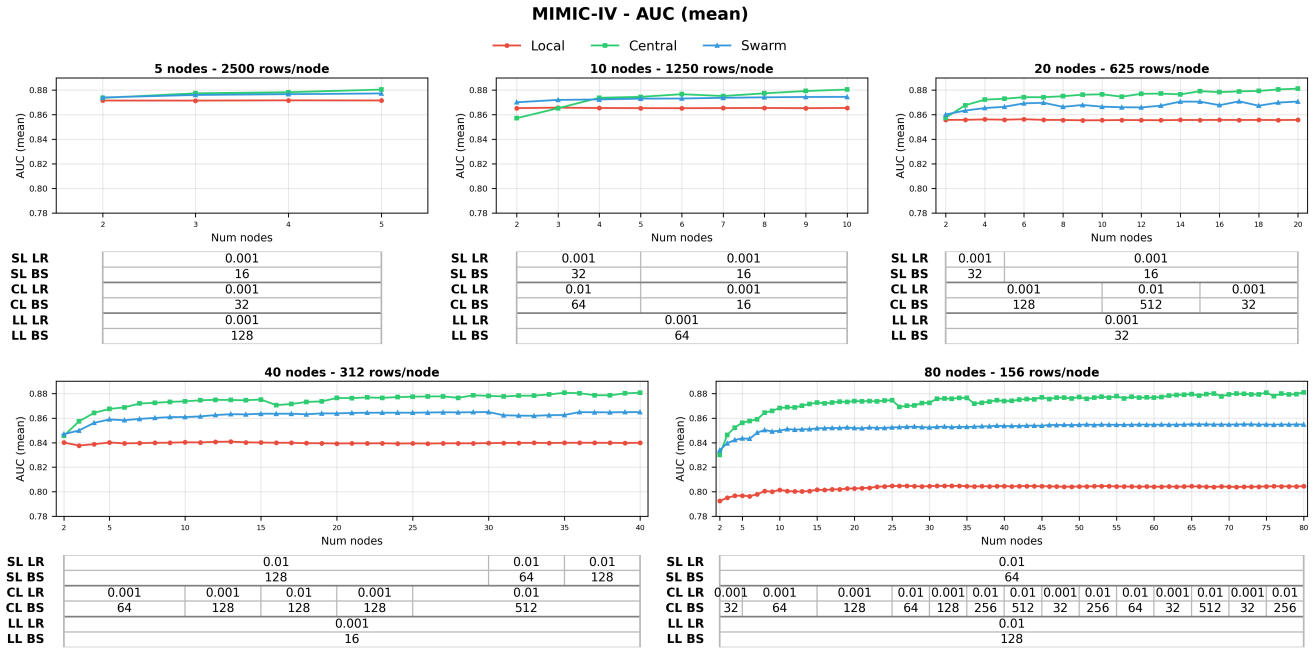


Figure 2: AUC comparison of swarm learning (SL), central learning (CL), and local learning (LL) across different experimental configurations for the *MIMIC-IV* dataset. Below each plot, we show the Batch Size (BS) and Learning Rate (LR) used for each experimental configuration, which were selected through a grid search every 5 nodes.

prevalence of 16.7%. In the *eICU* dataset, we were able to retrieve the required features from 25 hospitals, ranging from 1267 to 4670 admissions per hospital, with a total of 72 944 rows. The sepsis prevalence greatly varies across the different hospitals in *eICU*, ranging from 0.10% to 27.23%, with an average of 6.65%.

This extreme variance aligns with a structural characteristic of the *eICU* database: the diagnosis table is actively populated by clinicians during the ICU stay rather than being a retrospective administrative or billing record. Consequently, the broad range in sepsis prevalence may reflect not only potential epidemiological variations, but also heterogeneous clinical workflows, such as differing hospital monitoring protocols and sampling frequencies. While relying on these structured codes inherently introduces a degree of under-reporting, alternative approaches like retrospectively applying Sepsis-3 criteria would similarly encounter severe non-IID conditions due to the same heterogeneity in clinical workflows. Ultimately, this variance provides a highly realistic simulation of the label skew that decentralized models must handle in real-world collaborative networks.

To compare the results across the different datasets, we need to consider the constraints given by the number of rows or the number of nodes. To this extent, we created a version of *MIMIC-IV* with the same number of admissions as *MIMIC-III*, which is obtained by randomly sampling admissions from the original dataset with stratification. In *eICU*, instead, the nodes are represented by the hospitals, and therefore the maximum number of nodes is limited to 25. As a consequence, despite the large number of rows in *eICU*, many of these are removed to respect the constraints of the same number of patients per node.

4.3 Results and Discussion

In this section, we present and discuss the experimental results obtained from the *MIMIC-III*, *MIMIC-IV*, and *eICU* datasets. The analysis of the AUROC metric and the proposed KPIs provides insights into scalability, potential saturation points, and the impact of data granularity in swarm learning networks. All the experimental configurations described in Table 1 were executed, but for space reasons we show only the most significant ones.

4.3.1 AUC Comparisons. Before analyzing the KPIs, we evaluate the absolute predictive performance of the SL network using the AUROC curve. Figure 2 presents the results for the *MIMIC-IV* dataset, comparing the SL against the central and local learning models across different experimental configurations.

This comparison confirms an expected behavior: the central model acts as the upper bound, the local model as the lower bound, and the SL consistently positions itself between them. It can be observed that the SL performance suffers from high granularity, as the gap between the SL and the central model widens as the maximum number of nodes ($N_{\max}$) increases. Moreover, the central model continues to improve as more data is added, while the SL performance reaches a plateau and does not improve significantly after a relatively small number of nodes are added to the network. However, the SL model consistently outperforms the local models across all configurations, demonstrating that participating hospitals systematically benefit from joining the network.

It is important to emphasize that across all configurations (i.e., local, swarm, and central learning) we strictly maintained the exact same Feedforward Neural Network (FNN) architecture. Given our

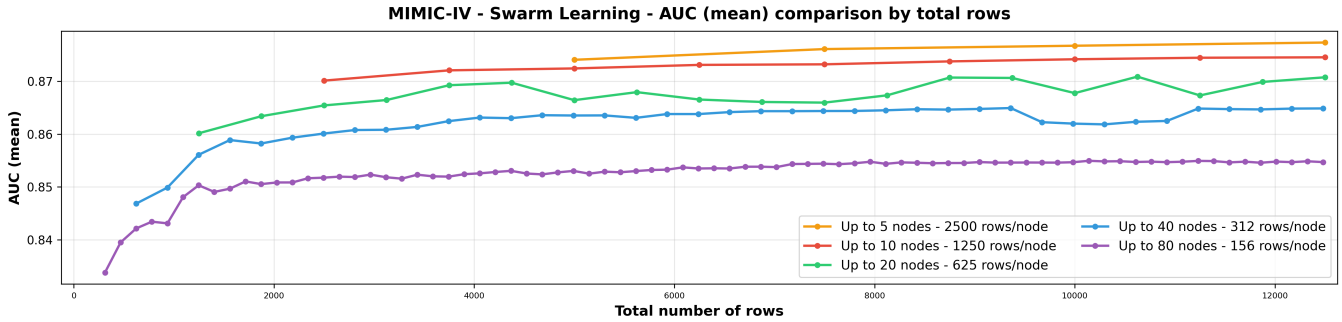


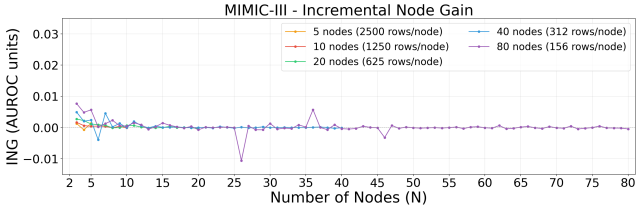
Figure 3: AUC comparison of SL networks across different experimental configurations for the *MIMIC-IV* dataset. Below each plot, we show the Batch Size (BS) and Learning Rate (LR) used for each experimental configuration, which were selected through a grid search every 5 nodes.

limited input space of exactly 41 tabular features and the highly granular setups evaluated (involving as few as 156 patients per node), we opted for this compact architecture to keep the model complexity proportional to the limited available data. By keeping this model constant, we isolate the effects of the distributed learning process and data granularity on predictive performance. The fact that the central baseline continues to improve its predictive performance with increasing data volumes demonstrates that the FNN has not reached its representational capacity limit. Therefore, the early plateau observed in the SL network is not caused by a model bottleneck, but is rather a limitation of the decentralized aggregation process and data fragmentation.

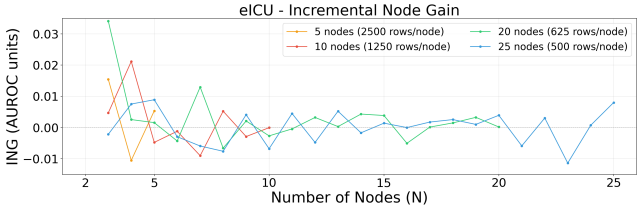
A more detailed analysis of the SL performance across different node configurations is provided in Figure 3, which shows a comparison of the SL performance based on the total volume of aggregated data. First of all, we can observe that the AUC is substantially higher as the number of patients per node increases, confirming that data granularity has a significant impact on the performance of the SL network. The saturation point is reached at the earlier stages of the network expansion for all the configurations, and this represents a clear limitation. In fact, when all the different configurations reach their maximum capacity ($N_{\max}$), the total amount of available data is identical across all setups (e.g., 12500 patients in *MIMIC-IV*), but the SL performance saturation prevents the highly granular networks (e.g., 40 or 80 nodes) to match the predictive performance of less granular setups (e.g., 5 or 10 nodes).

Overall, the absolute AUROC values demonstrate that data fragmentation limits the maximum achievable performance. The underlying dynamics of this limitation are formally quantified by the KPIs in the following subsections.

4.3.2 Incremental Node Gain (ING). As illustrated in Figure 4(a), the performance evolution in the IID scenario exhibits progressively smaller gains. The initial nodes integrated into the swarm provide a substantial improvement to the global model's AUROC, but this positive gain rapidly decreases toward zero. That is, regardless of the specific node configuration, the network efficiently extracts the majority of the available predictive information in the very early stages of collaboration. Conversely, Figure 4(b) reveals a high variability in the ING, where the addition of a node can result in a noticeable performance improvement or degradation, even at higher node counts. The ING gradually converges close to zero, but this instability visually demonstrates the algorithmic challenge of aggregating highly heterogeneous clinical data. In a non-IID setting, the localized data distributions across hospitals lead to conflicting weight updates, repeatedly perturbing the global model and preventing a smooth stabilization.



(a) *MIMIC-III* (IID)



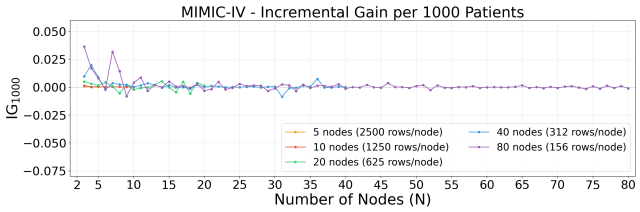
(b) *eICU* (Non-IID)

Figure 4: Incremental Node Gain for different experimental configurations in *MIMIC-III* and *eICU*.

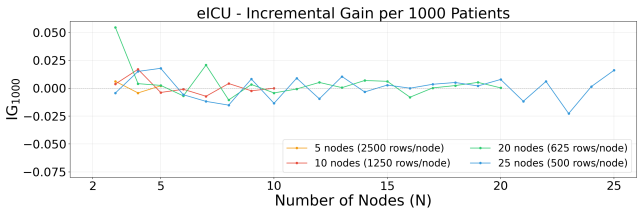
4.3.3 Incremental Gain per 1000 Patients (IG_{1000}). The analysis of the IID scenario (Figure 5a) demonstrates that dividing the same total amount of data into highly granular nodes (e.g., the purple curve for $N_{\max} = 80$, with only 156 patients per node) results in a highly variable learning contribution. In fact, the IG_{1000} values are higher with smaller node sizes, especially when the network is still small, but they all tend to rapidly decrease as more nodes are added until they all converge below the $IG_{1000} = 0.01$ threshold. This means that the prediction performance is obtained mainly at the initial stages, and the swarm network reaches a point where

the addition of a 1000 patients provides a negligible gain to the global model's AUROC, regardless of the specific granularity considered. In the non-IID scenario of *eICU* (Figure 5b), the IG_{1000} values are generally more unstable across all configurations, with a more pronounced negative impact for the higher granularities. The combination of high data granularity and heterogeneous hospital distributions creates more variability, and it is easily notable that the IG_{1000} values are often negative, especially for the higher granularities. This indicates that the heterogeneity of the data distributions across hospitals, combined with the small node sizes, leads to a more inefficient learning process.

The analysis of IG_{1000} reveals that high data granularity strictly penalizes the network's scalability. A fixed volume of clinical data returns better and more stable predictive gains when concentrated into a few large nodes rather than being dispersed across a highly granular swarm.



(a) *MIMIC-IV* (IID)



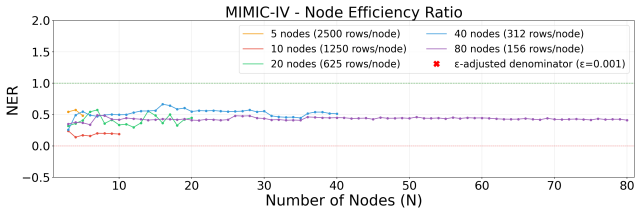
(b) *eICU* (Non-IID)

Figure 5: Incremental Gain per 1000 Patients (IG_{1000}) for different experimental configurations in *MIMIC-IV* and *eICU*.

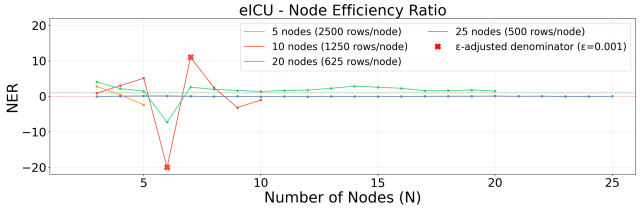
4.3.4 Node Efficiency Ratio (NER). To evaluate the swarm learning performance against a central learning approach, we analyze the Node Efficiency Ratio (NER). In the IID scenario (*MIMIC-IV*, Figure 6a), the NER stabilizes between 0.2 and 0.6. This behavior confirms that, although the swarm effectively improves its performance as more nodes are added, its predictive performance increases at a slower rate compared to the central model.

In highly heterogeneous environments, such as the non-IID *eICU* dataset (Figure 6b), adding new, differently distributed clinical data can cause the central model itself to reach local saturation or temporary degradation.

Specifically, in the 10-node configuration, this occurs at $N = 6$ and $N = 7$, where the central model incremental gain relative to $N = 2$ becomes extremely small ($\Delta_{\text{cent}}(6) \approx -4.4 \times 10^{-5}$ and $\Delta_{\text{cent}}(7) \approx 6.6 \times 10^{-4}$). In these cases, the ε-stabilized denominator prevents the NER from taking excessively out-of-scale values, and with $\varepsilon = 0.001$ it remains finite, although still large in magnitude (−20.0 and 11.0, respectively). These cases should therefore be interpreted with caution: they indicate that the central model provides almost no incremental gain, so the magnitude of the ratio is driven primarily by denominator saturation rather than by a substantial change in swarm performance. By contrast, later negative NER values (e.g., at $N = 9$ and $N = 10$) are not stabilization artifacts, but reflect a genuine divergence between swarm and central scaling behavior, with the central model falling below its $N = 2$ baseline while the swarm remains above it. Finally, let us note that in all the experimental configurations, only two cases of near-zero central incremental gain (i.e., $|\Delta_{\text{cent}}(N)| \leq \varepsilon$) are observed, and they are both present in Figure 6b.



(a) *MIMIC-IV* (IID)



(b) *eICU* (Non-IID)

Figure 6: Node Efficiency Ratio for different experimental configurations in *MIMIC-IV* and *eICU*.

5 Conclusions

In this paper, we presented a comprehensive scalability analysis of an SL network using both IID (i.e., *MIMIC-III* and *MIMIC-IV*) and non-IID (i.e., *eICU*) datasets. We systematically evaluated the performance evolution of the swarm network as new nodes are added, addressing key research questions related to scalability, saturation points, and the impact of data granularity. For this analysis, we introduced three novel Key Performance Indicators (ING, IG_{1000}, and NER) designed to quantify the learning outcome at the node, patient, swarm, and data levels.

Our experimental evaluation demonstrates that while isolated hospitals consistently benefit from joining the swarm network, its scalability is fundamentally limited by data granularity and distribution heterogeneity. Specifically, we observed that high data fragmentation (i.e., configurations with many small nodes) inherently limits the maximum achievable predictive performance. At a fixed volume of total available data, networks composed of a few large nodes achieve higher AUROC values compared to highly granular networks with many small nodes.

The analysis of the Incremental Node Gain (ING) demonstrates that the performance increments decrease progressively as the network expands. The SL network reaches a saturation plateau

before utilizing all available nodes ($N_{\max}$), and this saturation point is reached at a higher node count in highly granular setups. Also, the Node Efficiency Ratio (NER) confirms that while the SL network effectively benefits from additional data, its predictive performance improves at a slower rate compared to the central model.

Finally, the evaluation of the non-IID *eICU* dataset quantifies the impact of clinical data heterogeneity. The analysis of the incremental KPIs reveals that aggregating heterogeneous data introduces high variability in the performance evolution, often resulting in temporary performance degradation when new nodes are integrated.

Overall, these findings suggest that SL networks should favor the aggregation of larger local data partitions rather than high network fragmentation in order to optimize predictive efficiency.

In the future, we plan to extend this analysis by investigating different learning algorithms and aggregation strategies. Furthermore, to validate the generalizability of our findings, we aim to explore different ICU-related prediction tasks, such as mortality prediction, length of stay estimation, and disease onset prediction.

Acknowledgments

The research leading to these results has received funding from the European Union—NextGenerationEU through the Italian Ministry of University and Research under PNRR—M4C2-I1.3 Project PE_00000019 "HEAL ITALIA" to the authors CUP B33C22001030006. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

[1] Abien Fred Agarap. 2018. Deep Learning using Rectified Linear Units (ReLU). *CoRR* abs/1803.08375 (2018).

[2] Kallista Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Vladimir Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, et al. 2019. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems (MLSys)*, Vol. 1. mlsys.org, Stanford, CA, USA, 374–388.

[3] Trung Kien Dang, Xiang Lan, Jianshu Weng, and Mengling Feng. 2022. Federated learning for electronic health records. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 5 (2022), 1–17.

[4] R Phillip Dellinger, Mitchell M Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M Opal, Jonathan E Sevransky, Charles L Sprung, Ivor S Douglas, Roman Jaeschke, et al. 2013. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Critical care medicine* 41, 2 (2013), 580–637.

[5] Judith Sáinz-Pardo Díaz and Álvaro López García. 2023. Study of the performance and scalability of federated learning for medical imaging with intermittent clients. *Neurocomputing* 518 (2023), 142–154.

[6] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. 2021. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine* 49, 11 (2021), e1063–e1143.

[7] Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. Association for Computing Machinery, New York, NY, USA, 169–178.

[8] Stela Golovco, Matteo Mantovani, Carlo Combi, and John H. Holmes. 2022. Acute Kidney Injury Prediction with Gradient Boosting Decision Trees enriched with Temporal Features. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. 669–676. doi:10.1109/ICHI54592.2022.00135

[9] Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. 2019. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics* 99 (2019), 103291.

[10] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV. Version 2.2.

[11] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016), 1–9.

[12] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). OpenReview.net, San Diego, CA, USA, 1–15.

[14] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2017. Federated Learning: Strategies for Improving Communication Efficiency.

[15] Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, and Mosharaf Chowdhury. 2022. FedScale: Benchmarking Model and System Performance of Federated Learning at Scale. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, Baltimore, MD, USA, 11814–11827.

[16] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, Vol. 2. mlsys.org, Austin, TX, USA, 429–450.

[17] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. 2017. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Long Beach, CA, USA, 5330–5340.

[18] Matteo Mantovani, Carlo Combi, and Milos Hauskrecht. 2019. Mining compact predictive pattern sets using classification model. In *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*. Springer, 386–396.

[19] Matteo Mantovani, Carlo Combi, and Matteo Zeggiotti. 2019. Discovering and Analyzing Trend-Event Patterns on Clinical Data. In *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019, Xi'an, China, June 10-13, 2019*. IEEE, 1–10. doi:10.1109/ICHI.2019.8904774

[20] Matteo Mantovani, Riccardo Scheda, Carlo Combi, and Stefano Diciotti. 2024. A Key Performance Indicator to Analyze Swarm Learning Performances with EHR. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE, Orlando, FL, USA, 620–625.

[21] Anssi Nilsson, Samuel Smith, Gregor Ulm, Emil Gustavsson, and Mats Jirstrand. 2018. A performance evaluation of federated learning algorithms. In *Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning*. Association for Computing Machinery, New York, NY, USA, 1–8.

[22] Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data* 5, 1 (2018), 1–13.

[23] Oliver Lester Saldanha, Hannah Sophie Muti, Heike I Grabsch, Rupert Langer, Bastian Dislich, Meike Kohlruss, Gisela Keller, Marko van Treeck, Katherine Jane Hewitt, Fiona R Kolbinger, et al. 2023. Direct prediction of genetic aberrations from pathology images in gastric cancer with swarm learning. *Gastric Cancer* 26, 4 (2023), 541–553.

[24] Oliver Lester Saldanha, Philip Quirke, Nicholas P West, Jacqueline A James, Maurice B Loughrey, Heike I Grabsch, Manuel Salto-Tellez, Elizabeth Alwers, Didem Cifci, Narmin Ghaffari Laleh, et al. 2022. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nature Medicine* 28, 6 (2022), 1232–1239.

[25] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. 2021. Swarm learning for decentralized and confidential clinical machine learning. *Nature* 594 (2021), 265–270.

[26] Andrew C. Yao. 1982. Protocols for secure computations. In *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982)*. IEEE, Chicago, IL, USA, 160–164.

[27] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated learning with non-IID data.

[28] He Zhu, Jun Bai, Na Li, Xiaoxiao Li, Dianbo Liu, David L Buckeridge, and Yue Li. 2025. FedWeight: mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. *npj Digital Medicine* 8, 1 (2025), 286.