Article

A Federated Learning Approach to Support the Decision-Making Process for ICU Patients in a European Telemedicine Network

Giovanni Paragliola [1,*] , **Patrizia Ribino** [2] **and Zaib Ullah** [3]

1   Institute for High Performance Computing and Networking, National Research Council of Italy,
     80131 Naples, Italy
2   Institute for High Performance Computing and Networking, National Research Council of Italy,
     90146 Palermo, Italy; patrizia.ribino@icar.cnr.it
3   Università Telematica Giustino Fortunato, 82100 Benevento, Italy; zaib.ullah@unifortunato.it
*   Correspondence: giovanni.paragliola@icar.cnr.it

Abstract: A result of the pandemic is an urgent need for data collaborations that empower the clinical and scientific communities in responding to rapidly evolving global challenges. The ICU4Covid project joined research institutions, medical centers, and hospitals all around Europe in a telemedicine network for sharing capabilities, knowledge, and expertise distributed within the network. However, healthcare data sharing has ethical, regulatory, and legal complexities that pose several restrictions on their access and use. To mitigate this issue, the ICU4Covid project integrates a federated learning architecture, allowing distributed machine learning within a cross-institutional healthcare system without the data being transported or exposed outside their original location. This paper presents the federated learning approach to support the decision-making process for ICU patients in a European telemedicine network. The proposed approach was applied to the early identification of high-risk hypertensive patients. Experimental results show how the knowledge of every single node is spread within the federation, improving the ability of each node to make an early prediction of high-risk hypertensive patients. Moreover, a performance evaluation shows an accuracy and precision of over 90%, confirming a good performance of the FL approach as a prediction test. The FL approach can significantly support the decision-making process for ICU patients in distributed networks of federated healthcare organizations.

Keywords: federated learning; continuous learning; time series analysis; classification; healthcare informatics

1. Introduction

The SARS-CoV-2 pandemic highlights the need for better cooperation and knowledge sharing than in the past to prevent disease spread and maintain patient care quality. Notably, the pandemic showed that the uneven distribution of capacities and resources between healthcare organizations situated in small centers and those in urban areas means that the former are unable to provide the same quality of healthcare service as the latter. To address these challenges, a network of research institutions, medical centers, and hospitals all around Europe joined under the umbrella of the ICU4Covid project.

The ICU4Covid project [1], and its Cyber-Physical System for Telemedicine and Intensive Care (CPS4TIC), aims to create the sense of being part of the European telemedicine network and, in turn, to provide access to the network's capabilities, knowledge, and expertise.

Based on this mindset, the ICU4Covid project provides valuable ground for learning from real-world health data, which has proven to be effective in multiple healthcare applications, resulting in improved quality of care [2,3], prediction of disease risk factors [4,5], and analysis of genomic data for personalized medicine [6].

Nevertheless, a crucial aspect to be deliberated upon when operating within a healthcare ecosystem is the constrained accessibility and sharing of health data beyond the

primary institution due to the regulatory policies imposed by the EU General Data Protection Regulation (GDPR) [7]. Thus, traditional or centralized machine learning and deep reinforcement algorithms [8,9], which require aggregating such distributed data into a central repository to train a model, cannot be exploited. Leveraging such data while complying with data protection policies requires rethinking data analytics methods for healthcare applications. As a result, adopting decentralized artificial intelligence (AI) methods at the network edge plays a pivotal role in safeguarding privacy in healthcare applications.

Federated learning is a novel distributed interactive artificial intelligence paradigm that enables healthcare records that are located across different institutions to be connected without revealing personal information. The objective of this paper is to investigate the effectiveness of the FL approaches to exploit knowledge owned by each node of the ICU4Covid telemedicine network for AI model training while being compliant with the GDPR regulations for data privacy.

To achieve this aim, we integrated a federated learning architecture in each node of the CPS4TIC system. In doing so, the FL architecture enables the individual nodes of the network to act as local learners and send local model parameters to a central server instead of training data. The central server aggregates the local models, thus defining a single global model, which is sent back to the clients to proceed with the FL process until all rounds are completed. In FL, knowledge is exchanged by sending the models instead of the users' data, thereby preserving them.

The major contribution of this paper is a federated learning approach to support the decision-making process for ICU patients in a network of European telemedicine that provides the following advantages:

- Individual nodes independently train and collaboratively learn models without sharing local datasets.
- Model training takes place at an edge level that significantly facilitates organizations that operate with data privacy as their top priority.
- The FL approach ensures lower latency than the centralized cloud-based model training approach.
- Finally, the FL approach allows for balancing data intelligence. With it, small and medium-sized healthcare organizations can benefit from collective intelligence without requiring large datasets.

To validate the proposed approach to support decision making in a network of federated healthcare organizations (i.e., hospitals, clinics, etc.), we also show how the knowledge owned by a single organization is spread among all the federation members by improving the reliability of the local model as a prediction test. Finally, we evaluated the FL algorithms using the most common metrics in the field of machine learning, such as precision, recall, and accuracy. Performance evaluation shows an accuracy and precision of over 0.91, confirming a good performance of the FL approach as a prediction test.

In addition, a quantitative and qualitative estimation of improvement caused by the federated process over the local nodes is provided, reporting an enhancement of the performance of up to 38%.

The rest of the paper is organized as follows. Section 2 provides a brief background about FL and its application in healthcare. Section 3 introduces the working context in which the proposed FL approach is developed. Sections 4–6 present the proposed federated learning approach, its application as a predictive test, and its performance evaluation, respectively. Finally, conclusions are drawn in Section 7.

2. Related Work

2.1. Background of Federated Learning

FL is a decentralized, collaborative ML methodology first coined by Google in an era when users' data privacy became a high priority. FL presents an ML environment where multiple entities work together to solve an ML problem coordinated by a central server. The raw data remain local to each client and are not transferred or exchanged. Updates

focused on aggregation are employed to achieve the learning goal [10]. Recent research endeavors are focused on utilizing FL for secure data privacy. The existing literature is more oriented toward exploring FL on distributed devices, e.g., mobile user interactions, communication cost optimization, server reliability, uneven data distribution for model training, and many more [11].

To define an FL problem, suppose $i = 1, \ldots, N$ represents the set of N participants, where each trains its own local AI model with its distinct dataset D_i. FL aims to fine-tune the weight parameters w of the global model such that the loss function values for all local AI models are minimized. Hence, each participant trains an ML model locally and communicates the resulting model parameters to the centralized FL server. Consequently, the central FL server aggregates all participants' local models to construct a global model. The global model is again shared with all participants. The process is repeated round-by-round until a pre-defined accuracy is achieved.

2.2. Application of FL in Healthcare

This section provides the most recent research developments involving FL applications in telemedicine and decision making of smart healthcare.

2.2.1. Applications in Telemedicine

In this subsection, we provide a brief overview of the most recent research endeavors regarding FL techniques' role in telemedicine from data privacy and decision-making perspectives. The authors in [12] constructed a mechanism to utilize the FL approach in e-health for data security and latency. The proposed FL mechanism's effectiveness was proven over traditional data security techniques. Chen et al. [13] attempted to improve the privacy aspect of patients' Electronic Health Records (EHRs) by countering irrelevant and unnecessary update issues affecting the model's predictability. The proposed architecture, called PFL-IU, is consistent with irrelevant updates and enhances model convergence and the degree of prediction accuracy. Lamia et al. [14] explored the role of FL techniques in IoMT to counter the COVID-19 pandemic. The proposed framework consists of fog layer utilization within the healthcare system. The fog layer is responsible for data pre-processing and is used for model training to investigate COVID-19 using FL. Vijay et al. [15] proposed an FL-based protocol to collectively design classifiers and calibrate various IoT devices in the healthcare industry. A linear calibration model is used to improve the gain and offset deviations. Training occurs at the individual device level, and resultant classifier parameters are shared to learn and construct a global classifier model. Han et al. [16] proposed an FL-based zero-watermarking framework to handle teledermatology data privacy and security. In the proposed architecture, the FL trains a sparse autoencoder network to retrieve dermatology image features. Furthermore, the extracted image features are transformed into a two-dimensional Discrete Cosine Transform to generate zero watermarking. Jinshan et al. [17] proposed a joint architecture of FL and edge computing to safeguard patients' data privacy and security. Initially, a lightweight framework called KubeFL was used to assist in health monitoring and diagnosis. In the second phase, an FL-based training model is used over device–edge–cloud layering with the highest accuracy of 95.8%. Wang et al. [18] proposed a joint architecture consisting of the blockchain, federated learning-BFGF, and genome-wide association studies to ensure individual data security. The Automated Quality Control implementation ensures the quality of training data before training models at local levels. Blockchain eliminates malevolent elements to ensure users' security. Simulation of the proposed architecture shows that it can better ensure the security of genetic data. Aich et al. [19] proposed an architecture comprised of blockchain and FL techniques to ensure the real-time safety of patient data in the local and global healthcare system. Lucian et al. [20] proposed FL-based COVID-19 detection using pre-trained deep learning models. The framework efficiency was tested using three different local databases without sharing them. The mechanism successfully detected medical images thought to be suggestive of COVID-19 patients. Durga et al. [21] developed an innovative architecture

consisting of FL and blockchain technology to detect COVID-19. The FL ensures model complexity optimization while the blockchain maintains the decentralization of datasets and privacy. The architecture collects data from various medical centers' databases, develops a learned model, predicts COVID-19 with sufficient accuracy, and maintains patients' data privacy by only allowing access to authorized people.

2.2.2. Application in Decision-Support System

In this subsection, we briefly explore the research efforts and utilization of FL techniques in healthcare and telemedicine. Linardos et al. [22] utilized FL on cardiovascular magnetic resonance data attained from various datasets. The authors employed a pre-trained 3D-CNN network while using four data augmentation arrangements and thoroughly observing their effects on various learning choices. The FL-based framework proved more robust and attained competitive performance compared to the centralized learning models. Chuhan et al. [23] presented a communication-efficient FL technique called FedKD, employing dynamic gradient compression mechanisms. FedKD optimizes communication costs by up to 94.89% and attains competitive performance regarding centralized model learning. Nidal et al. [24] proposed a framework to control the pandemic while effectively preserving users' privacy. The architecture comprises Unmanned Aerial Vehicles (UAVs) serving users' equipment and a proof-of-concept network. Using ambient sensors, a lightweight FL model deployed on UAVs collaboratively learns COVID-19 symptoms attained from user equipment (e.g., cell phones). The resultant COVID-19 data are shared with the relevant authority for effective utilization. Nguyen et al. [25] proposed a decentralized FL approach employing a knowledge distillation mechanism to ensure the privacy and security of medical data. Each server executes independently at a local level without needing external factors' interference. The proposed model performance is superior to that of the traditional centralized training model approach.

2.2.3. Comparison with the Literature

The previous sections provide a broad picture of the application of FL in healthcare, ranging from applications that address COVID-19 to those guaranteeing data privacy in decision-support systems. All cited works show the benefits of adopting FL technologies to their use case. However, it is possible to highlight the main differences between the proposed paper and related works as follows:

(1) *Use Case Application*. To the best of our knowledge, we present the first federated learning study on ICU decision-support systems and demonstrate that FL performance is comparable to that of traditional approaches while preserving patient privacy.
(2) *Privacy/Performance Trade-off Definition*. The cited papers do not address the trade-off between privacy and performance; instead, we provide a first formulation and implementation considering the experiment results and traditional centralized models.
(3) *Explicit Validation Scenario*. All cited papers lack experimental validation; instead, we offer it and illustrate how the knowledge of the clients merged in a single aggregated model from a temporal point of view, reporting a quantitative and qualitative estimation of the improvements achieved by the training process.
(4) *Data Type*. Previous works show that only a few analyzed papers address the problem of time-series classification. Our use-case application, to the best of our knowledge, is the first application to the training of a time-series-based identification model in a federated environment.

These aspects represent the gaps in the current knowledge and highlight the novel contribution of this paper, which makes it different from the state-of-the-art. Table 1 summarizes these aspects.

Table 1. Comparison between related work and the proposed work.

Paper	Use Case	Trade-Off Privacy/Performance	Validation	Data Type
[12]	EHR	×	×	Time series
[13]	EHR	×	×	Numerical
[15]	Photoplethysmography	×	×	Numerical
[16]	Teledermatology	×	×	Image
[17]	Health Monitoring	×	×	Numerical
[18]	Genome-wide Association	×	×	Numerical (generated by software)
[20]	COVID-19 Detection	×	×	Image
[22]	Diagnosis Disease	×	×	Image
[23]	News Recommendation	√	×	Text
[24]	COVID-19 Detection	×	×	Image
[25]	Prediction Viability Embryo	√	×	Image
Proposed Paper	Hypertension Early Detection	√	√	Time series

3. Working Context

3.1. Overview of ICU4COVID European Project

The pandemic showed that centralized ICUs' monitoring and their uneven distribution of capacities between the rural and urban areas remains a big challenge in effectively countering the spread of COVID-19. Hence, real-time information sharing and cooperation between hospitals, healthcare workers, and the public became significant in containing COVID-19 and ensuring high-quality healthcare services. The Cyber-Physical System for Telemedicine and Intensive Care (CPS4TIC) aims to use technology to help hospitals share information and deliver care, especially in rural areas. This system reduces the risk of infection for healthcare workers. CPS4TIC includes a telemedicine cockpit, a console in each hospital, a connectivity platform, and smart bedside hubs; see a conceptual view in Figure 1. The ICU4Covid project aims to make CPS4TIC work in many European hospitals. ICU4Covid helps to test and deploy CPS4TIC on a larger scale with healthcare professionals and hospitals. This technology allows separate intensive care units to be turned into a connected system, where a central hospital is connected to its smaller hospitals in a region.

Typically, an ICU hub includes a main ICU and affiliated smaller hospitals. They use telemedicine and monitoring tools to help caregivers assess, diagnose, and treat patients in the ICU. Each ICU hub is equipped with advanced technology such as 5G, radar sensors, and NVIDIA AI chips. Figure 1 shows a group of connected ICU hubs in an area. These hubs are managed by a central control station called the Integration Centre, where AI models are tested and deployed, decisions are made, and knowledge is shared. The ICU4Covid plan ensures that both existing and new ICUs work together as one ICU system. It does not depend on the hospital's setup and keeps telemedicine safe.
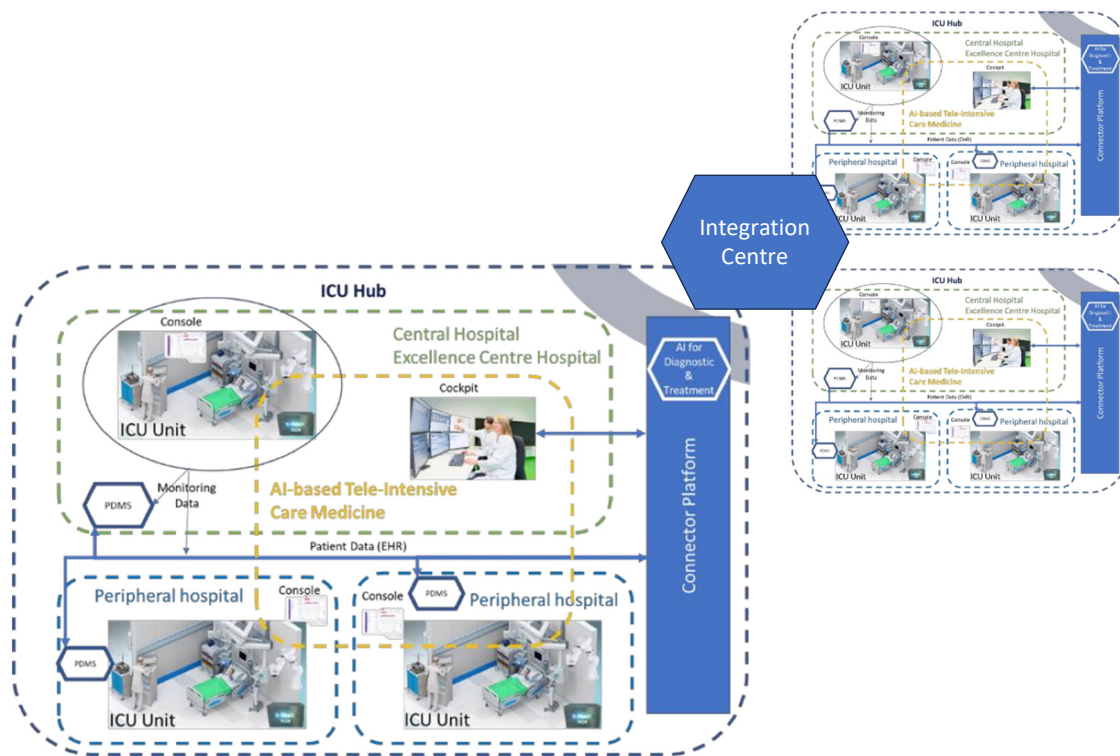
Figure 1. Interconnected ICU hubs' architecture across a region or zone (extracted from [1]).

3.2. CPS4TIC Enhanced with Federated Learning

The ICU4Covid project aims to support the decision-making process for ICU patients in a network of European telemedicine providers by defining predictive models to assist health professionals in the best patient treatment; therefore, collecting and processing all data are core tasks. Since ICU4Covid involves different health infrastructures physically located across various hospitals, moving data between all nodes raises potential privacy and security risks. To guarantee knowledge sharing between each ICU hub, we have proposed and integrated a federated learning (FL) architecture in each node of the CPS4TIC system. Figure 2 shows a logical view of the proposed integration between the as-is CPS4TIC system and the FL architecture for the experiments.
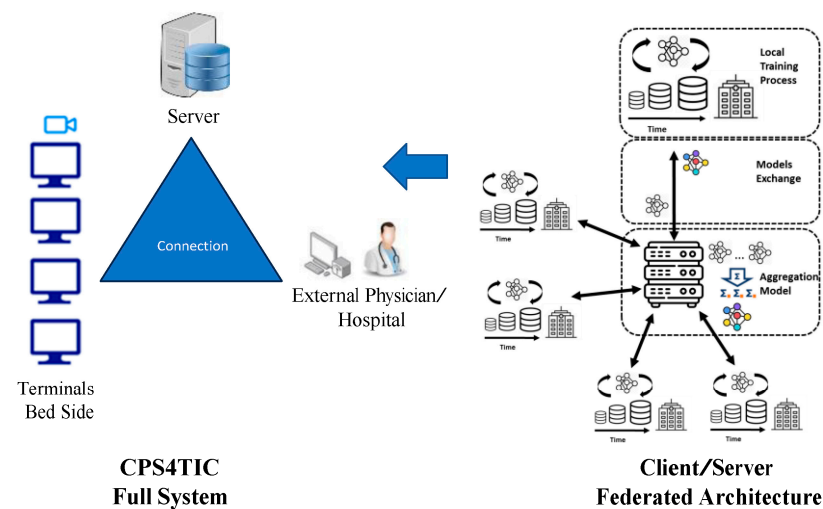

Figure 2. Integration of federated learning and CPS4TIC.

The left side of Figure 2 provides a comprehensive view of the architecture inherent to each ICU hub. In this depiction, a localized central server orchestrates all regional data sources and then catalogues this information in a dedicated local PostgreSQL database.

To seamlessly incorporate the federated learning (FL) architecture, the right side of Figure 2 comes into focus. Within this delineation, each central server assumes an extended configuration that includes a client node equipped with the ability autonomously to refine deep learning models. Conversely, the role of the aggregator server materializes within the domain of the integration center nodes, strategically located outside the boundaries of the interconnected networks of the CPS4TIC ICU hub. It is worth noting that the training process is defined to run parallel with normal clinical activities.

4. Materials and Method

4.1. Federated Learning Scenario

This section introduces the FL architecture adopted for the ICU4Covid project and the scenario under analysis. In this scenario illustrated in Figure 3, the FL architecture consists of three clients respectively associated with *Hospital 1*, *Hospital 2*, and *Hospital 3*, with the same data structure provided by the CPS4TIC hosted in each hospital. The participant's data remain within the hospital that owned them, and they are used for local training. Individually, each participant participates in training a localized neural network, using their respective local datasets to develop a unique local model. Following this process, updates to these models contributed by each participant are sent to a central aggregator server. This aggregator server acts as a computational hub that holds the trust of each participating participant. The disparate local models are merged within the aggregator's domain to form a cohesive global consensus model. This consolidated global model is then redistributed to each participant, facilitating an additional round of localized training. Participants establish communication channels with the aggregator server for connectivity purposes using remote procedure calls mediated by a secure transport layer security network connection. An encrypted channel securely exchanges information between the participants and the aggregator server to ensure the confidentiality and integrity of sensitive data—including the model itself—optimizer weights, and aggregated metrics.
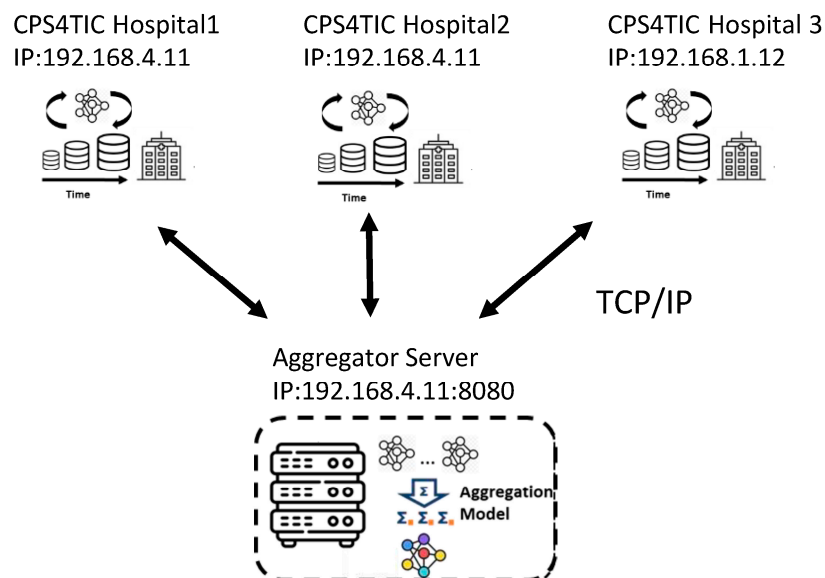

Figure 3. Logic overview of the benchmark validation scenario.

For the sake of experimentation, these three participants are emulated on local clusters. To build the system, the following set of operational hypotheses about the nodes was considered: (1) the structure of the neural network remains consistent across nodes; (2) the temporal aspects associated with uploading and downloading the local model are assumed

to be negligible; (3) each node is assumed to have a satisfactory level of security; and (4) hardware capabilities are assumed to be uniform across all nodes.

4.2. Dataset

To validate the proposed FL approach, the SHAREE [26] database was used to emulate health data owned by different hospitals. In the scenario under study, such data are used to train a machine learning model for the early identification of subjects at a higher risk of developing fatal cardiovascular events. In so doing, each hospital that takes part in this federation, using real-time data from patients hospitalized in the ICUs, will exploit the knowledge acquired by all hospital members of the federation to improve its ability to make appropriate healthcare decisions and provide better healthcare services.

The SHAREE database stores 169 electrocardiographic (ECG) registrations of hypertensive patients monitored for 24 h with a halter to record major cardiovascular and cerebrovascular events. The recordings of patients who had experienced a dangerous event were labeled *high-risk*, while the remaining were marked *low-risk*. Hence, to emulate three hospitals, such a database is split into three datasets to prove that a participant with fewer cases of hypertensive patients will have the same reliability in identifying high-risk patients as a participant with higher registered cases. The data used for the training process are a multivariate time series (MTS), where the raw signal is described by three metrics, V1, V2, and V3, describing three electrodes placed on the subject's chest during the monitoring. Figure 4 shows an example of a sample recorded in the dataset.
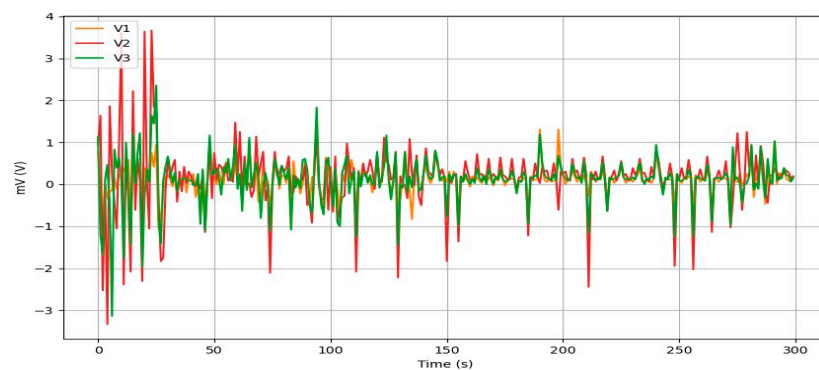

Figure 4. Generic signal sample for the model validation.

Portions of 5 min are randomly extracted from the input data, with each portion defining a training sample. Each second of each training sample models a one-time point. Thus, all training samples have 300 time points, and each time point is a vector of spatiotemporal features. The training set is defined as follows:

$$x = [(V_1, V_2, V_3)_1, (V_1, V_2, V_3)_2, \ldots, (V_1, V_2, V_3)N], \forall x \in TS$$

where $N = 300$ and each feature is characterized by V_1 (mv), V_2 (mv), and V_3 (mv).

4.3. FL Algorithm

Federated learning is a distributed machine learning algorithm that builds a global model by averaging weights ω across many participants over several communication rounds t. Each neuron in the hidden layer has a transfer function that takes each feature in a sample and multiplies it by its weight. Such weights are the learnable parameter and are modified during the training of the neural networks in each client. This section illustrates the algorithms that implement federated learning from both the client and server sides. Algorithm 1 presents the federated learning process from the server's point of view. This process essentially consists of two basic phases: initialization and learning.

In the initialization phase, the server assigns initial values to parameters intended to be shared among the clients. These parameters include the weight, denoted ω_0, and the

learning rate, indicated as η_0. In addition, the server sets a threshold, denoted by *Th*, which is used to evaluate the accuracy of the local models.

During each iteration of the learning phase, the server accumulates the model parameters of each client. Subsequently, clients that have achieved a local accuracy above the specified threshold *Th* are considered eligible for inclusion in the *Candidate Set*. It is important to note that during this phase, a subset of clients—totaling *m*—is randomly selected from the *Candidate Set* to participate in the aggregation process.

In this paper, *m* is set to half of the *Candidate Set* size. In the following equation:

$$w_{t+1} = \sum_{k=1}^{K} (n_k * n) * w_{t+1}^k$$

ω_{t+1} describes the aggregation strategy adopted by the server, where n_k is the sample size of the *k*th client and n is the total number of training samples. ω_{t+1} comes from the *k*th client in round *t* + 1. Finally, the aggregation weight ω_{t+1} is spread to all clients.

Conversely, Algorithm 2 shows the FL process from the client's perspective.

In this context, each client performs an individualized local training procedure using the updated parameters obtained from the server. At the end of the training iteration, the model is evaluated to determine whether the achieved accuracy exceeds a predefined threshold *Th*. This evaluation serves as a criterion for determining whether the model is suitable for transmission to the server node. If the local model's accuracy exceeds the threshold *Th*, the newly updated local model parameters are sent to the aggregation server (Algorithm 1), (Algorithm 2).

Algorithm 1: Test-Based FedAvg ServerSide

---

1 Run on the server
2 Initialize ω_0, *Th* and η;
3 foreach *round t = 0,1,...,T* **do**
4   **foreach** *client k = 1,2,...,K in parallel* **do**
5 $(\omega_{t+1}^k, Acc_k) \leftarrow ClientUpdate(k, \omega_t, \eta)$;
6 $CandidateSet_t \leftarrow \{client_k \mid Acc_k > Th\}$;
7 $S_t \leftarrow Random(CandidateSet_t, m)$;
8   **foreach** *client k in S_t* **do**
9   Retrieve ω_{t+1}^k from client *k*;
10   **foreach** *client k in S_t* **do**
11 $\omega_{t+1} \leftarrow \sum_{k=1}^{K}(n_k/n) * \omega_{t+1}^k$;
12   Spread ω_{t+1} to Clients;

Algorithm 2: Test-Based FedAvg: ClientSide

---

1 Run on each *k* local client
2 Collect ω from the server;
3 Update Local Model;
4 foreach *local epoch e=1,2,..,E* **do**
5 $\omega \leftarrow BackPropagation(\omega)$;
6   *Acc* ⤺ Evaluate ω;
7 if *Acc > Th* **then**
8   return (ω, Acc) to the server;
9 else
10   return (Acc) to server;

4.4. Hybrid Neural Network

In the context of early detection of high-risk hypertension, traditional window-by-window approaches have limitations. They only consider the temporal order of the entire sequence during training, and this lack of temporal order information is due to the random batch processing of samples during training sessions.

Hence, we propose a solution that exploits the power of a Long Short-Term Memory (LSTM) network, which treats the entire time series as a unified entity. This approach preserves the temporal order and ensures accurate classification. What distinguishes our neural network from traditional methods is not the input structure, which remains a time series, but the unique way our network processes and analyzes these data.

Figure 5 gives an overview of a hybrid network, which means it comprises different deep learning networks.

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 300, 20)	1920
conv1d (Conv1D)	(None, 300, 8)	1608
max_pooling1d (MaxPooling1D)	(None, 30, 8)	0
conv1d_1 (Conv1D)	(None, 30, 4)	164
max_pooling1d_1 (MaxPooling1	(None, 3, 4)	0
flatten (Flatten)	(None, 12)	0
dropout (Dropout)	(None, 12)	0
dense (Dense)	(None, 2048)	26624
dropout_1 (Dropout)	(None, 2048)	0
batch_normalization (BatchNo	(None, 2048)	8192
dense_1 (Dense)	(None, 512)	1049088
batch_normalization_1 (Batch	(None, 512)	2048
dense_2 (Dense)	(None, 256)	131328
dense_3 (Dense)	(None, 128)	32896
batch_normalization_2 (Batch	(None, 128)	512
dense_4 (Dense)	(None, 1)	129
activation (Activation)	(None, 1)	0

Total params: 1,254,509

Figure 5. Summary of the network architecture.

The model starts at the top level with an input layer that receives a multivariate time series (MTS) as a sample. The layers labeled *lstm1* and *conv1d* combine stacked convolutional and recurrent networks. These networks automatically extract important features from the raw data signals. Next, the layers named *denso1* to *denso7* consist of seven densely connected layers. These layers make a decision by classifying the input as either high risk or low risk. The specific values used for training, which are a subset of the chosen hyper-parameters, are specified for each layer.

Concerning building the optimum network structure and tuning the hyper-parameters, a grid search approach was adopted to optimize the setting of the values, such as the number of neurons per layer, the learning rate, and the number of dense layers. The implementation of the network was achieved by adopting the TensorFlow framework.

5. Application of the Proposed Approach as a Prediction Test to Support Decision Making

To assess the proposed approach as a reliable predictive test to support decision making in a network of federated hospitals, such as provided by the ICU4Covid, this section shows how the knowledge owned by a single hospital is shared within the federation.

A patient's electrocardiogram (ECG) sample has been extracted from the CPS4TIC dataset of client 1 (CPS4TIC-1). This particular sample is categorized in the *high-risk* class. The patient associated with this sample, referred to as *Patient$_x$*, has his or her data stored exclusively within CPS4TIC-1. Consequently, CPS4TIC-2 and CPS4TIC-3 do not gain any specific insight from this patient's data. Figure 6 shows the validation of the proposed approach in the context of the ICU4Covid project. At the beginning of the federated training process, only CPS4TIC-1 accurately classifies the validation sample. Conversely, CPS4TIC-and CPS4TIC-3 fail. The aggregated model also struggles with classification, as our experimental results indicate the need for additional rounds to effectively aggregate knowledge from the local models. In the middle stages of the FL process, CPS4TIC-1 continues to accurately classify the validation sample. Remarkably, the aggregated model is also able to achieve this classification. This result signifies the successful integration of knowledge from the CPS4TIC-1 model into the aggregated model through the federated learning process.
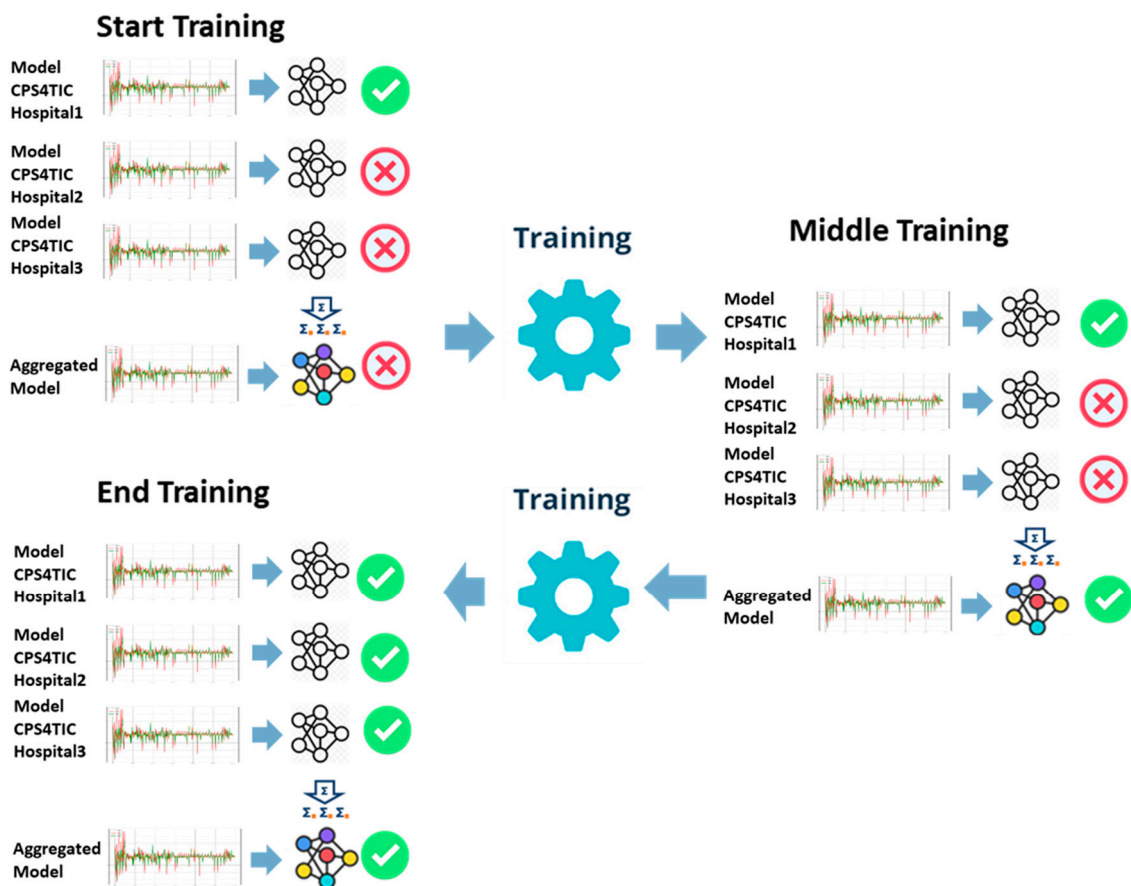

Figure 6. Logical overview of the validation of the federated process.

Figure 7 presents a qualitative view of the accuracy trend throughout the training process in each local node and for the aggregated model. The term *Start Training* corresponds to the first round of communication within the federated learning process. *Middle Training* refers to the halfway point of the learning process, which occurs after 10 out of a total of 20 rounds. Finally, *End Training* is the final communication round. It is possible to note the changes in accuracy following the stage of the training, confirming what is shown in Figure 6.
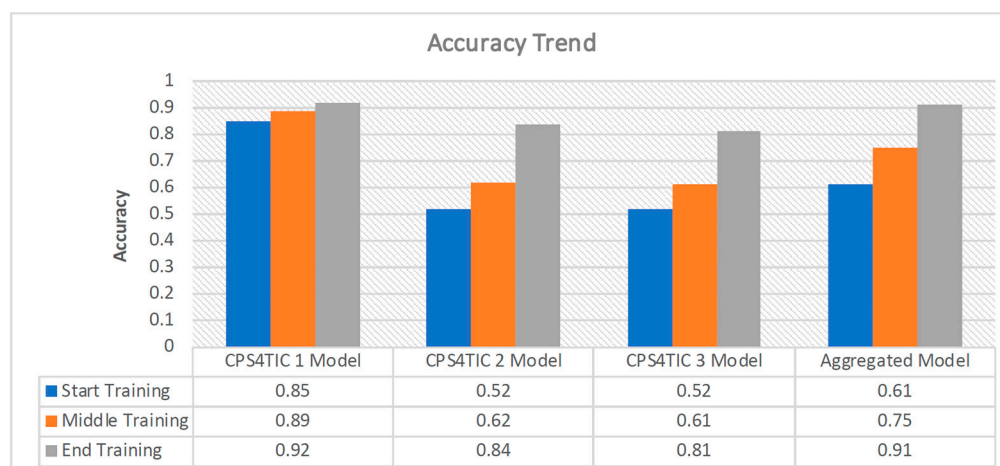
Figure 7. Accuracy trend for the local models and the aggregated model.

As can be seen from Figure 7, locally within CPS4TIC 1, accuracy is maintained at consistently high values from the beginning of training. In contrast, the other nodes and the aggregated model show accuracy values of 52% and 61%, respectively.

As the rounds progress, CPS4TIC 1 and the aggregator model show noticeable performance improvements, reaching accuracy of 89% and 75%, respectively. This aspect is consistent with the federated learning (FL) algorithm, where the insights of the aggregator model are diffused to CPS4TIC 2 and CPS4TIC 3. This results in significant performance improvements by the end of training, with CPS4TIC 2 and CPS4TIC 3 reaching accuracy of 84% and 81%, respectively.

At the end of the federated learning process, all models demonstrate the ability to accurately classify the validation sample. This result is consistent with the proposed approach since the server progressively exchanges the refined parameters from the clients during each round. Consequently, at the culmination of the process, all clients encapsulate the combined knowledge of all nodes.

This result is significant. Without a federated learning process, Hospital 2 and Hospital 3 may have categorized $Patient_x$ as a low-risk individual, potentially resulting in inappropriate care protocols for the patient. However, by integrating knowledge from the aggregated model, $Patient_x$ can be correctly identified as high risk. This accurate classification enables healthcare professionals at Hospital 2 and Hospital 3 to formulate and implement appropriate healthcare measures for the patient's well-being.

The accuracy trend in Figure 7 highlights the improvement resulting from the federated process for all local models. In particular, CPS4TIC 2 and CPS4TIC 3 show an enhancement from the beginning to the end of the training of 38% and 35%, respectively.

It is worth highlighting that hospitals in small regions, on average, have much fewer cases than hospitals in large urban areas; thus, their knowledge of particular cases can be limited. Conversely, the probability that hospitals in large urban areas manage unusual cases is higher, thus resulting in an improved quality of care provided to their patients. Hence, another advantage of using a federated learning approach in a distributed network of hospitals is improving small hospitals' knowledge in classifying cases the hospital has never treated.

6. Experimental Results and Discussions

This section presents the experimental results concerning the performance evaluation of the proposed FL approach achieved within the ICU4Covid project.

6.1. Training and Testing Configurations

The training set consists of more than 14,000 samples, evenly distributed between the two classes, *High-Risk* and *Low-Risk*. Each client has its test set which is defined using the

hold-out approach. It consists of 700 samples, two-thirds labeled as low-risk and one-third as high-risk. The total number of samples in the training set was equally split across the three local nodes, (i.e., 4746). Each local node i splits its portion of the local dataset D_i so that the training set and the validation set are equal to 90% (i.e., 4000 samples) and 10% (i.e., 474 samples) of D_i. The validation set adopted a 10-fold cross-validation loop that was performed for the model optimization and to avoid over-fitting. In addition, from the original dataset (SHAREE [26]), a global test dataset was created to provide a comprehensive comparison between all local clients, federated aggregate, and centralized models at the end of the training process. The global test dataset contains a total of 177 samples.

Table 2 summarizes the total training data in each local node.

Table 2. Review of the training data in each node.

Node	Full Set	Training Set	Validation Set	Local Test Set
CPS4TIC-1	4746	4000	474	700
CPS4TIC-2	4746	4000	474	700
CPS4TIC-3	4746	4000	474	700

For the experiments, all clients were emulated on local clusters with 73 biprocessor 16 Core nodes with 512/1024 GB of RAM and a 1.2 TB disk, 292 NVIDIA V100 16/32 GB GPUs, InfiniBand network at 100 Gb/s, and elastic storage having 1.3 PB of raw capacity.

6.2. Performance Metrics

The pertinent parameters for determining performance metrics for machine learning models encompass the four components within the confusion matrix as follows:

$$CM = \begin{bmatrix} TP & FN \\ FP & FN \end{bmatrix} \tag{1}$$

where *True Positive* (*TP*) indicates the number of positive samples that have been correctly classified. *True Negative* (*TN*) indicates the number of negative samples that have been correctly classified. *False Positive* (*FP*) indicates the number of samples that have been incorrectly classified as positive. *False Negative* (*FN*) indicates the number of samples that have been incorrectly classified as negative.

To evaluate the results of this paper, we used the most common metrics derived from the confusion matrix: accuracy, precision, *F*1-score, specificity, sensitivity (recall), and Matthews's correlation coefficient (MCC), defined as follows [27,28].

Precision (PREC) is the ratio between correctly classified samples and all samples assigned to that class.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (REC), also known as *Sensitivity*, indicates the rate of positive samples correctly classified and is computed as the ratio between correctly classified positive samples and all samples assigned to the positive class.

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{3}$$

Specificity (SPEC) is the negative class version of the recall and denotes the rate of negative samples correctly classified.

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

The *F1-score* (*F1*) is the harmonic mean of precision and recall, penalizing extreme values of either.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Cecall} \tag{5}$$

Accuracy (ACC) is the ratio between the number of samples that are correctly classified and the overall number of samples in the dataset being assessed. This metric is widely utilized for machine learning applications in the field of medicine, although it is recognized for its potential misrepresentation when confronted with varying class proportions. This metric is extensively employed in machine learning applications in the healthcare domain; however, it is acknowledged for its potential to be misleading for datasets with dissimilar class proportions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Finally, the *Matthews correlation coefficient* (MCC) can be considered a measurement of correlation between the actual and predicted classifications. MCC achieves high values only if the classifier obtains good results in all the entries of the confusion matrix. Unlike the previous metrics, which range within [0, 1], *MCC* has the range of [1, 1], where a value of 1 represents a perfect prediction, 0 is equivalent to random guessing, and −1 represents a complete discordance between the predicted outcomes and the actual observations. MCC is popular in ML settings because of its favorable properties in cases of imbalanced classes where ACC can be misleading.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (FN + FN)}} \tag{7}$$

6.3. Performance Results

This section analyzes the performance of the federated and centralized scenarios to assess the proposed approach's effectiveness. Specifically, in the *federated scenario*, the proposed ML model for predicting the risk of cardiovascular events is implemented according to the federated configuration reported in Figure 3. Conversely, in the *centralized scenario*, the same model is used in a traditional centralized configuration. The meaning of these results concerns the definition of a baseline to assess the goodness of the federated results.

Table 3 shows the confusion matrix of both federated and centralized scenarios. The table was defined by evaluating both centralized and federated models over the global test set. As we can see, the predictive model obtained by adopting a federated approach is able to detect 117 and 42 as true instances and 13 and 5 as false instances. On the other hand, the centralized scenario is able to detect 120 and 53 as true instances and 4 as false instances.

Table 3. Confusion matrix related to federated and centralized scenarios.

Scenario	TP	TN	FP	FN
Federated	117	42	13	5
Centralized	120	53	2	2

According to this matrix, Figure 8 shows the values of performance computed according to Equations (2)–(7). It is worth noting that concerning the *federated scenario*, all metrics came from the average over the evaluation of the local models over the global test set.

The *centralized scenario* achieved accuracy, sensitivity, and precision values of 98% and specificity of 96%. The *F1-score* reached 98%. These metrics confirm the goodness of a hybrid neural network model for early detection of high-risk hypertension. On the other hand, the *federated scenario* achieves an accuracy and a precision of 90%, a sensitivity of 96%, and a specificity of 76%, while the *F1-score* reached 93%.
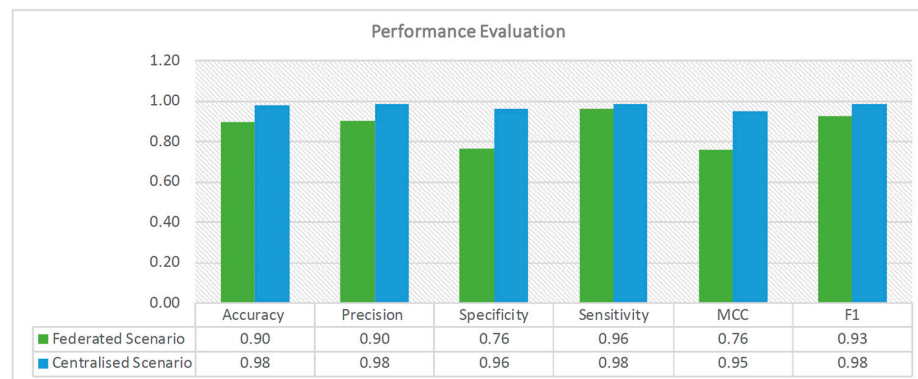
Figure 8. Performance comparison between federated and centralized scenarios.

As expected, the *centralized scenario* performed best. However, the federated scenario reached high values of accuracy, precision, and sensitivity, which are among the most important metrics when working in the healthcare domain, where it is fundamental to recognize positive diagnoses. On the contrary, the specificity decreased by 19.9% with respect to the centralized scenario.

Moreover, as also expected, the MCC is higher in the centralized scenario than in the federated one. Indeed, MCC, as previously mentioned, is a statistical metric that produces a high score only if the prediction obtains good results in all of the four confusion matrix categories, proportionally both to the size of positive elements and the size of negative elements in the dataset, thus providing a balanced measure of the model efficiency. The centralized scenario confirms the model's goodness in working with an unbalanced dataset. However, in the federated scenario, the value of MCC decreases to 0.76, compared to 0.95 of the centralized one. It is worth pointing out that the FL approach is conceived to work with more than three nodes, and thus has the possibility of exploiting a higher collaboration among nodes than in this small scenario. Consequently, MCC and other metrics may reach better values than the previous one.

Moving deeper into the evaluation of the federated approach, Figure 9 shows the performance in terms of accuracy achieved by each local model on the global test set at the end of the learning process. It is worth noting that the test set is the same for each model since its purpose is to fairly compare all the models. Figure 9 shows how the aggregated model performs slightly better than the local models. Indeed, the accuracy of the aggregated model is equal to 90%, which is higher than that of each local model, of 87%, 88%, and 88%, respectively. It is worth noting that the results shown in Figure 9 refer to the end of the federated learning process after all rounds have been addressed.
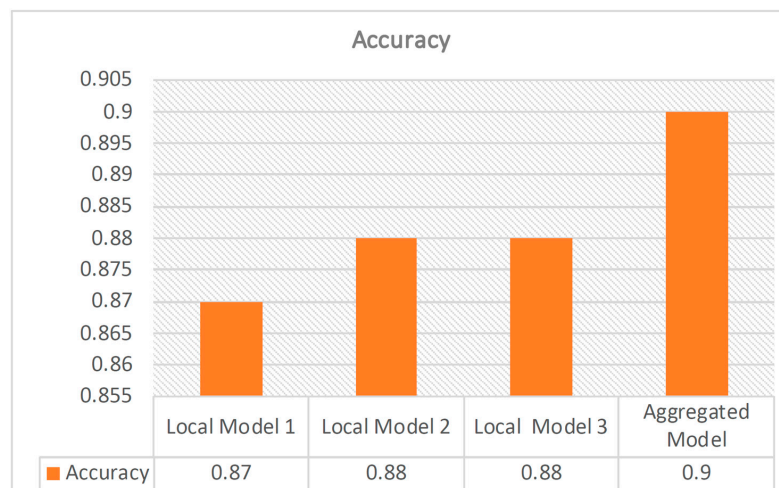

Figure 9. Accuracy comparison between each local model and the aggregated model.

Figure 10 shows the trend of the accuracy of the aggregated model during the rounds. It is possible to note that during the beginning rounds, the model exhibited low accuracy, confirming that the process needs more rounds to correctly merge the knowledge inside all local models. Halfway through the federated learning process, the accuracy stabilized at over 0.8, and reached the final value of 0.90 at the end of the process.
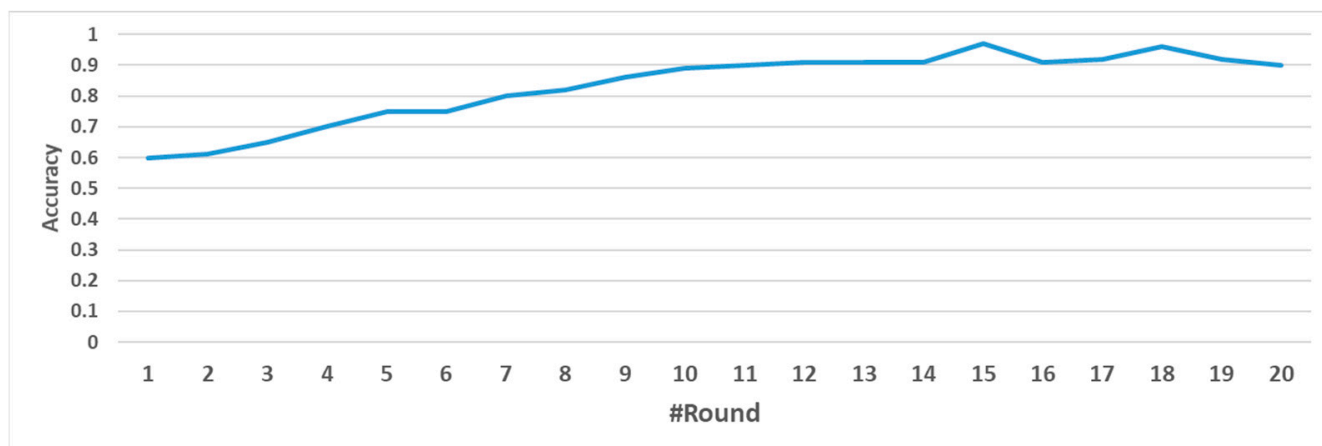

Figure 10. Trend of the accuracy of the aggregated model during the training.

All these results indicate that federated learning is a promising approach in a telemedicine network for exploiting local knowledge on the network nodes, guaranteeing data privacy by accepting only a small loss in performance when compared with centralized approaches.

6.4. Comparison with Traditional Machine Learning Approaches

To evaluate the efficacy of the proposed federated learning approach compared to conventional methods, a comparative analysis was conducted by assessing our results against performance measurements obtained from six traditional classifiers operating on the same dataset. These classifiers are the best ones found in the original work of Melillo et al. [29] and include the boosting meta-learning approach (AB), C4.5 decision tree, artificial neural networks using a multilayer perceptron (MLP), naïve Bayes classifier (NB), random forest (RF), and support vector machine (SVM).

The bar chart depicted in Figure 11 shows the performance outcomes of the aforementioned classifiers and the federated classifier. The statistical measures of accuracy, sensitivity, and specificity were obtained from the study conducted by Melillo et al. [29]. Notably, these values have been rounded to the second decimal place. As we can see, the federated approach obtains better performance in terms of accuracy with respect to traditional approaches for predicting cardiovascular events. In contrast, specificity is less compared to that of the majority of the other approaches, whereas the sensitivity of 96% is a better result than that achieved by the traditional classifiers. In particular, the sensitivity of the proposed federated scenario is almost doubled with respect to MLP, and it also shows a considerable gain with respect to the other approaches, which reach the highest value of 71%. Sensitivity, as previously mentioned, is regarded as being among the most important metrics for medical studies since it is desirable to miss as few positive instances as possible. Thus, the proposed federated approach can be considered a good choice for implementing AI-based applications in the healthcare domain.
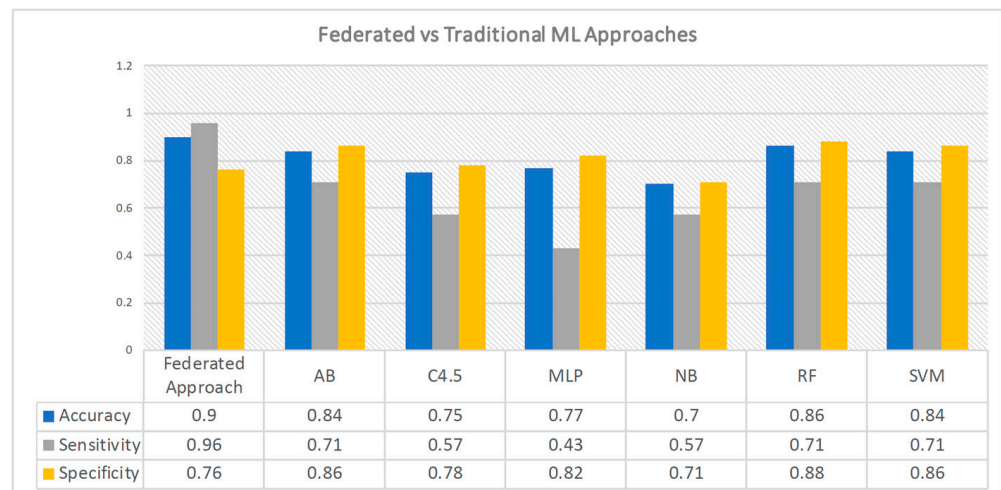
Figure 11. Comparison with traditional machine learning approaches.

6.5. Discussion

The performance results achieved by the federated model can be considered satisfactory since the overall results reach a good level in terms of accuracy and other metrics. They also have a higher degree of cost effectiveness than the centralized mode, which guarantees superior performance in classification tasks. However, it must be emphasized that in the centralized mode, all data must be transferred from their respective storage nodes to the designated node responsible for the learning process. Consequently, this undertaking introduces an inherent element of risk concerning data security and privacy.

The centralized approach clearly necessitates the transfer of 100% of the data across nodes, consolidating it in a single central node. In contrast, the federated approach bypasses the transfer of even a single raw data byte. Instead, it involves the transfer of federated views, effectively mitigating any potential privacy risks associated with handling the data. For the sake of the truth, we have to underline that the federated approach can introduce communication cost issues while sharing the neural network parameters; these issues were evaluated and addressed in [29,30].

Our results highlight a trade-off between performance and security, shown in Figure 12, which visually describes the relationship between these two aspects for a generic predictive model. The figure defines a qualitative visual space delimited by the projection of privacy and performance. The figure does not define quantitative values since the purpose is only to provide a simple and immediate tool for centralized and federated model comparison.
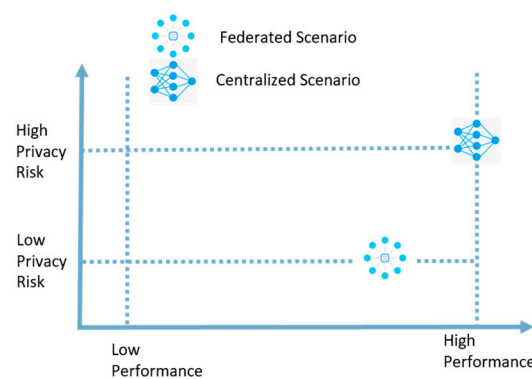

Figure 12. Graphical view of trade-off between performance and security.

The upper-right quadrant illustrates this trade-off in a conventional predictive model based on a centralized approach, characterized by robust performance metrics but with high exposure to privacy risks. In contrast, within the federated perspective, performance levels,

while commendable, do not rise to the heights of centralized models but are accompanied by significantly lower, nearly negligible, privacy risk. Our research aims at the lower right quadrant, where performance metrics approach those of centralized models while maintaining minimal privacy risk.

Another important consideration is the training time. As shown in Figure 8, the federated model shows commendable results in terms of classification performance when compared to the centralized model. However, the temporal disparity must be emphasized; the aggregated federated model requires a longer time commitment of 3 to 7 days, as opposed to the 7 to 24 h required by the centralized model. This temporal discrepancy is due to the need to go through numerous rounds to achieve convergence in the aggregated federated model.

7. Conclusions

The paper has presented a federated learning approach to support the decision-making process for ICU patients in a network of European telemedicine. A network of federated healthcare organizations evaluates the goodness of the results to show how the knowledge of every single organization spreads among all the federation members by improving the reliability of the local model as a prediction test. Finally, traditional metrics quantify the performance of the proposed approach compared with the centralized one.

The results confirm that the FL approach can significantly support the decision-making process for ICU patients in distributed networks of federated healthcare organizations.

The proposed approach's main limitation is the global model's decreasing performance as the number of clients increases. As shown in Figure 9, the local models have a lower accuracy level than the global models. While these results may appear acceptable when dealing with a few clients, they become increasingly unfavorable when the number of clients exceeds ten.

To ensure a more robust validation of the approach, it is imperative to increase the number of clients involved in the training process and to use a larger dataset that includes heterogeneous data information.

Future works will address the issues faced during the experiments concerning the number and trustworthiness of clients and dataset dimensions, such as:

(1) Evaluation of the impact of new clients: It is essential to evaluate the solution with a more significant number of clients and, in addition, define a way to establish the trustworthiness of a new client involved in the initial training process.
(2) Evaluation of a larger dataset: The second issue will assess the learning process with a larger dataset.

Author Contributions: Conceptualization G.P.; methodology, G.P. and P.R.; software, G.P. and Z.U.; validation, G.P. and Z.U.; formal analysis, G.P. and P.R.; investigation, G.P.; resources, Z.U.; data curation, Z.U.; writing—original draft preparation, G.P. and P.R.; writing—review and editing, G.P., Z.U. and P.R.; visualization, Z.U.; supervision, G.P. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: https://physionet.org/content/shareedb/1.0.0/, accessed on 23 January 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. ICU4Covid-Cyber-Physical Intensive Care Medical System for COVID-19. Available online: https://www.icu4covid.eu/ (accessed on 23 January 2023).
2. Johnson, A.E.; Ghassemi, M.M.; Nemati, S.; Niehaus, K.E.; Clifton, D.A.; Clifford, G.D. Machine learning and decision support in critical care. *Proc. IEEE* **2016**, *104*, 444–466. [CrossRef] [PubMed]
3. Chakrabarty, A.; Zavitsanou, S.; Sowrirajan, T.; Doyle, F.J., III; Dassau, E. Getting IoT-ready: The face of next-generation artificial pancreas systems. In *The Artificial Pancreas*; Academic Press: Cambridge, MA, USA, 2019; pp. 29–57.

4. Poplin, R.; Varadarajan, A.V.; Blumer, K.; Liu, Y.; McConnell, M.V.; Corrado, G.S.; Webster, D.R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2018**, *2*, 158–164. [CrossRef] [PubMed]

5. Zheng, T.; Xie, W.; Xu, L.; He, X.; Zhang, Y.; You, M.; Yang, G.; Chen, Y. A machine learning-based framework to identify type 2 diabetes through electronic health recordsInternational. *J. Med. Inform.* **2017**, *97*, 120–127. [CrossRef] [PubMed]

6. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef] [PubMed]

7. Voigt, P.; Von dem Bussche, A. *The EU General Data Protection Regulation (GDPR). A Practical Guide*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2017; Volume 10, pp. 10–55.

8. Naeem, M.; Coronato, A.; Ullah, Z.; Bashir, S.; Paragliola, G. Optimal user scheduling in multi antenna system using multi agent reinforcement learning. *Sensors* **2022**, *22*, 8278. [CrossRef] [PubMed]

9. Ismail, A.; Fiorino, M.; Abbas, M.; Syed, M.H.; Ullah, Z. Cloud-Based Monitoring System for Personalized Home Medication. In Workshop Proceedings of the 19th International Conference on Intelligent Environments (IE2023), Island, Mauritius, 27–30 June 2023; IOS Press: Amsterdam, The Netherlands, 2023; pp. 113–124.

10. Konečný, J.; McMahan, H.B.; Ramage, D.; Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv* **2016**, arXiv:1610.02527.

11. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **2019**, *10*, 1–19. [CrossRef]

12. Sellamna, A.; Boukhamla, A.Z.; Benkaddour, M.K. Federated Learning paradigm in E-health systems: An overview. In Proceedings of the 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS), Oum El Bouaghi, Algeria, 12–13 October 2022; pp. 1–5.

13. Chen, H.; Li, H.; Xu, G.; Zhang, Y.; Luo, X. Achieving privacy-preserving federated learning with irrelevant updates over e-health applications. In Proceedings of the ICC 2020–2020 IEEE International Conference on Communications (ICC), Virtual, 7–11 June 2020; pp. 1–6.

14. Fourati, L.C.; Samiha, A.Y. Federated learning toward data preprocessing: COVID-19 context. In Proceedings of the 2021 IEEE International Conference on Communications Workshops (ICC Workshops), Virtual, 14–23 June 2021; pp. 1–6.

15. Anavangot, V.; Lukose, J. Calibration and Classifier Design for IoT Healthcare Applications using Federated Learning. In Proceedings of the 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), Kochi, India, 23–25 June 2022; pp. 1–6.

16. Han, B.; Jhaveri, R.; Wang, H.; Qiao, D.; Du, J. Application of robust zero-watermarking scheme based on federated learning for securing the healthcare data. *IEEE J. Biomed. Health Inform.* **2021**, *27*, 804–813. [CrossRef] [PubMed]

17. Lai, J.; Song, X.; Wang, R.; Li, X. Edge intelligent collaborative privacy protection solution for smart medical. *Cyber Secur. Appl.* **2023**, *1*, 100010. [CrossRef]

18. Wang, H.; Zhang, X.; Xia, Y.; Wu, X. An intelligent blockchain-based access control framework with federated learning for genome-wide association studies. *Comput. Stand. Interfaces* **2023**, *84*, 103694. [CrossRef]

19. Aich, S.; Sinai, N.K.; Kumar, S.; Ali, M.; Choi, Y.R.; Joo, M.I.; Kim, H.C. Protecting personal healthcare record using blockchain & federated learning technologies. In Proceedings of the 2022 24th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 13–16 February 2022; pp. 109–112.

20. Florescu, L.M.; Streba, C.T.; Șerbănescu, M.S.; Mămuleanu, M.; Florescu, D.N.; Teică, R.V.; Nica, R.E.; Gheonea, I.A. Federated Learning Approach with Pre-Trained Deep Learning Models for COVID-19 Detection from Unsegmented CT images. *Life* **2022**, *12*, 958. [CrossRef] [PubMed]

21. Durga, R.; Poovammal, E. FLED-Block: Federated Learning Ensembled Deep Learning Blockchain Model for COVID-19 Prediction. *Front. Public Health* **2022**, *10*, 892499. [CrossRef] [PubMed]

22. Linardos, A.; Kushibar, K.; Walsh, S.; Gkontra, P.; Lekadir, K. Federated learning for multi-center imaging diagnostics: A simulation study in cardiovascular disease. *Sci. Rep.* **2022**, *12*, 3551. [CrossRef] [PubMed]

23. Wu, C.; Wu, F.; Lyu, L.; Huang, Y.; Xie, X. Communication-efficient federated learning via knowledge distillation. *Nat. Commun.* **2022**, *13*, 2032. [CrossRef]

24. Nasser, N.; Fadlullah, Z.M.; Fouda, M.M.; Ali, A.; Imran, M. A lightweight federated learning based privacy preserving B5G pandemic response network using unmanned aerial vehicles: A proof-of-concept. *Comput. Netw.* **2022**, *205*, 108672. [CrossRef] [PubMed]

25. Nguyen, T.V.; Dakka, M.A.; Diakiw, S.M.; VerMilyea, M.D.; Perugini, M.; Hall, J.M.; Perugini, D. A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data. *Sci. Rep.* **2022**, *12*, 8888. [CrossRef] [PubMed]

26. Melillo, P.; Izzo, R.; Orrico, A.; Scala, P.; Attanasio, M.; Mirra, M.; De Luca, N.; Pecchia, L. Automatic prediction of cardiovascular and cerebrovascular events using Heart Rate Variability analysis. *PLoS ONE* **2015**, *10*, e0118504. Available online: http://journals.plos.org/plosone/article?id=10.1371635 (accessed on 23 January 2023).

27. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061. [CrossRef]

28. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. In *BMC Genomics BMC Genomics (Volume 21, Issue 1)*; Springer Science and Business Media LLC: Berlin, Germany, 2020; p. 645.

29. Paragliola, G. Evaluation of the trade-off between performance and communication costs in federated learning scenario. *Future Gener. Comput. Syst.* **2022**, *136*, 282–293. [CrossRef]

30. Paragliola, G.; Coronato, A. Definition of a novel federated learning approach to reduce communication costs. *Expert Syst. Appl.* **2022**, *189*, 116109. [CrossRef]