Client Recruitment for Federated Learning in ICU Length of Stay Prediction

Vincent Scheltjens*† *IEEE Graduate Student Member*, Lyse Naomi Wamba Momo*,
Wouter Verbeke† *IEEE Member*, Bart De Moor* *IEEE Fellow*
*KU Leuven, Department of Electrical Engineering (ESAT),
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics,
Kasteelpark Arenberg 10, 3001 Leuven, Belgium.
†KU Leuven, Faculty of Economics and Business,
Naamsestraat 69, 3000 Leuven, Belgium.

***Abstract*—Machine and deep learning methods for medical and healthcare applications have shown significant progress and performance improvement in recent years. These methods require vast amounts of training data which are available in the medical sector, albeit decentralized. Medical institutions generate vast amounts of data for which sharing and centralizing remains a challenge as the result of data and privacy regulations. Federated Learning (FL) is well-suited to tackle these challenges. However, FL comes with a new set of open problems related to communication overhead, efficient parameter aggregation, client selection strategies and more. In this work, we address the step prior to the initiation of a federated network for model training, client recruitment. By intelligently recruiting clients, communication overhead and overall cost of training can be reduced without sacrificing predictive performance. Client recruitment aims at pre-excluding potential clients from partaking in the federation based on a set of criteria indicative of their eventual contributions to the federation. In this work, we propose a client recruitment approach using only the output distribution and sample size at the client site. We show how a subset of clients can be recruited without sacrificing model performance whilst significantly improving computation time. By applying the recruitment approach to the training of federated models for accurate patient Length of Stay prediction using data from 189 intensive care units (clients), we show how the models trained in federations made up from only recruited clients significantly outperform federated models trained with the standard procedure in terms of predictive power and training time.**

***Index Terms*—Federated Learning, Client Recruitment, Length of Stay**

I. Introduction

Recent machine learning (ML) and deep learning (DL) techniques have proven to be of significant value for healthcare applications [2], [3], [25]. An abundance of medical data is continuously generated, especially in Intensive Care Units (ICU) where patients are monitored uninterrupted and vitals are charted continuously. DL techniques are particularly good at extracting underlying complex relations in such large datasets, thus the vast amount of available data bodes well for these data hungry models. One of the main limitations in relation to medical data of any sort relates to data availability. As a result of regulatory restrictions both in Europe through the General Data Protection Regulation (GDPR) [7] and more recently the European Union Artificial Intelligence Act, and the United States with the Health Insurance Portability and Accountability Act (HIPAA) [4] imposing similar restrictions, data sharing poses a significant challenge in the medical sector both for primary and secondary use [6]. In both cases, sharing and centrally accumulating privacy-sensitive data is bounded by a legal framework, necessary to protect the privacy of the individuals behind the data. This, however, poses a challenge for research and applications that require large amounts of medical data.

In recent years, the value of federated learning (FL) has been illustrated for medical applications, especially in the field of computer vision for medical image segmentation and classification [13], [21], [27]. Here, the FL approach allows to learn complex models over decentralized data without direct access to the data. FL, however, comes with a new set of challenges. In the real world, decentralized data is often not independent and identically distributed *(non-IID)* which harms the training procedure and model performance [33]. The reduced performance can mainly be attributed to the weight divergence in local models as a result of the *non-IID* data [33].

Selection procedures of federated training algorithms may sample those contributors whose data is most informative at training time. This, however, requires all potential contributors to form part of the federation and participate in at least one round of training. In the work by Yichen Ruan et al. [24], the theoretical foundations are provided for a client recruitment process which precedes the initiation of the federation. This aims at building the federation with those clients for which it can, a priori, be said that their contributions to the federation will be valuable. More specifically, clients for which the local

This work was supported by KU Leuven: Research Fund (projects C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316), Industrial Research Fund (Fellowships 13-0260, IOFm/16/004, IOFm/20/002) and several Leuven Research and Development bilateral industrial projects; Flemish Government Agencies: FWO: EOS Project no G0F6718N (SeLMA), SBO project S005319N, Infrastructure project I013218N, TBM Project T001919N; PhD Grants (SB/1SA1319N, SB/1S93918, SB/1S1319N), EWI: the Flanders AI Research Program VLAIO: CSBO (HBC.2021.0076) Baekeland PhD (HBC.20192204) and Innovation mandate (HBC.2019.2209) European Commission: European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC Adv. Grant grant agreement No 885682); Other funding: Foundation 'Kom op tegen Kanker', CM (Christelijke Mutualiteit)

data does not form a good representation of the population are considered less representative and therefore, of less value to the federation. Intuitively, when aiming for accurate global predictive performance, less representative clients result in higher weight divergence in the local models at training time [33], which harms the training procedure and predictive power of the resulting global model.

Whilst in [24], a sound optimization framework is provided for client recruitment, in this work, by building on [24], we define a client recruitment approach and the construct of client-level representativeness using only the local target distribution and sample size. Subsequently, the practical relevance of this approach is demonstrated through an application on the real-world eICU dataset [8], [19], [20] for which only 54 out of the 189 potential clients are recruited. With the recruited client subset making up the federation, we are able to learn better performing models, in terms of predictive power and training time, to predict patient Length of Stay (LoS) in ICU compared to the standard FL approach as proposed in [15].

The remainder of this work is structured as follows. Section II outlines the related work. In Section III the basics of FL are discussed. In Section IV the client recruitment approach is discussed along with the experimental setup and corresponding description of the data. Section V outlines the results and provides a discussion of the main findings. In Section VI concluding remarks are presented followed by a discussion of the current limitations and directions for future work.

The repository containing the code and instructions for reproducing the results is available on GitHub here: github.com/vscheltjens/eicu-cl-recr.

II. Related Work

Federated learning was originally proposed by McMahan et al. in 2017 [15] in conjunction with the *FedAvg* algorithm and has since been proven to be a valuable learning framework that can yield accurate models without direct access to the local data.

A significant amount of research efforts have been dedicated to the client selection problem. Client selection in FL deals with client scheduling for each round of training, i.e., for each training round a subset of clients is selected that will contribute in the next training iteration. The standard approach is to randomly select this subset. However, as argued in multiple studies [17], [29], [30], a better approach is to impose criteria based upon which client selection can be performed. These criteria often relate to how informative updates from certain clients are, or the local computational resources [31], [32]. This, however, does not allow for noninformative clients to be pre-excluded from the federation, which is where client recruitment comes in. Client recruitment aims to discard potential clients from the set of available clients for the federation, prior to initiation of the latter. One of the ways to do so is by considering limited statistics on the local dataset that do not contain privacy sensitive information. By pre-excluding potential clients, the foremost benefit is that the overall cost of training a model in the federated setting is reduced [24] without sacrificing predictive performance. In addition, optimally, the least informative clients are pre-excluded which results in increased predictive performance for the resulting global model.

Although FL originates in the large-scale distributed edge computing setting, it has been extensively studied in healthcare [13], [16], [18], [21]. Specifically the work discussed in [16] closely relates to the work proposed here. In [16] the authors extensively assess different parameter settings for federated training on medical ICU data to identify suitable parameters for In Hospital Mortality (IHM) binary classification. In this study, we tackle a different problem, i.e., ICU LoS prediction, using the same eICU dataset. The authors from [16] show that a larger number of local training epochs improves performance whilst reducing the training cost at once as a result of the reduced server-client communication. In this work we extend upon these previous works by proposing a client recruitment scheme using limited statistics on the local data, applied to the critical care setting.

III. Federated Learning Basics

Federated Learning (FL) [15] is a technique that allows for a central model to be trained over distributed data that originate and are stored locally. This approach was originally proposed within the setting of large-scale distributed learning on mobile devices and has been widely used and researched in the medical sector given the privacy enhancing aspect as well as the regulatory restrictions on data sharing within the sector. Each local data source, e.g., a mobile phone or a server hosted at a hospital, is referred to as a *client*. The group of clients that contribute to training a central model, including the central *server* is referred to as the *federation*. The central server orchestrates the learning process over the different clients following the predefined FL algorithm. In that sense, the central server is holistically responsible for (i) initiating the model; (ii) providing a copy of the model to all of the clients; (iii) aggregating over the received model parameters and (iv) send back the updated model to the clients. Each client locally trains the model for a predetermined number of epochs after which only the model parameters are sent back to the central server. In the standard *FedAvg* algorithm [15] considered in this work, local model parameters are averaged into the global model update. In addition, for each communication round, either all or a subset of the clients in the federation are randomly selected for training. We consider this the standard FL setting which comes with most out of the box implementations.

In this work, the clients correspond to 189 hospitals where the main challenge relates to some extent to the concept of *client shift* in conjunction with the understanding that in the real world, data from different hospitals is likely to be *non-IID* distributed. In Fig. 2, 4 of the in total 189 local target distributions are shown, illustrating how even when only looking at the target, the data hosted at each of the clients can vastly differ from one client to the next. More specifically, clients host data that originates from different

hospitals in different geographic regions, each with potentially different demographic characteristics, etc.. Depending on local sample size and training parameters, models may overfit to the local data and bias the global model when aggregating model parameters. These are common problems to FL training algorithms and have been tackled by introducing more involved aggregation algorithms such as *Weighted FedAvg* or smart client selection [17], [29], [30].

Whilst FL allows to learn a central model from distributed medical data in a privacy enhancing manner, not all data at each of the hospitals are equally valuable. More specifically, smaller local sample sizes are less likely to represent the global distribution which will typically result in a larger empirical training loss, as discussed in [24]. In addition, as a result of different local demographics, even larger local datasets could also diverge in distribution from the population, which will again reduce predictive performance. Existing client selection and aggregation algorithms provide partial solutions to these issues. These methods, however, require for all clients to form part of the federation and for each client to have participated at least once in a training round. If not, the algorithms do not obtain the required information to guide the further aggregation and selection procedure.

This work builds on the client recruitment work introduced in [24] and the field of FL in healthcare to develop a method for client recruitment by looking only at the local distribution of the target variable and the local sample size. With this, the recruitment process aims to recruit clients for which the local data better represents the population before initiating the federation. This would intuitively lead to better performing models globally and reduced training time.

IV. Methods

To assess the performance and utility of federated models that are trained with a subset of clients recruited following the approach that will be described in IV-C, a single prediction task is defined for both the central, and federated models. The task is for each model to predict the patient LoS in ICU, similar to what has been studied in [1], [22], [23], [28] where LoS is defined as the remaining time in ICU for a given patient. Formally, the task is to yield predictions $\hat{y}$ which approximate y, the true LoS.

A. Data

For the evaluation of the proposed approach on real-world, multi-center data, the eICU dataset [8], [19], [20] is used, covering over 200,000 eICU admissions to 208 US hospitals. The total admission count covers over 139,000 unique patients registered at one of the US hospitals between 2014 and 2015.

The dataset is made publicly available for research purposes and is particularly interesting for FL as it allows for the data to be mapped to the originating institution. The data is preprocessed in concordance with the preprocessing pipeline proposed in [23]. In summary, the first 24 hours of data post ICU admission, for adult patients are extracted and used to predict LoS. Both temporal and static information are extracted, cleaned, re-sampled, imputed and one-hot encoded. The temporal data is fused with the static patient data which will serve as the input to the models. For patients with multiple recorded stays over the designated time period, only one of the stays is considered to avoid information leakage when obtaining train, test and validation splits. We refer to the work by Rocheteau et al. [23] for an extensive description of the preprocessing pipeline.

The resulting cohort is comprised of data pertaining to 89,127 patient stays over 189 hospitals, covering a total of 35 features of which 17 are temporal and 18 static. The summary statistics for the resulting data cohort are included in Table I and a detailed overview is provided in Appendix A.

TABLE I
Overview of the extracted and preprocessed data cohort

Number of patient stays	89,127
Train	62,375
Validation	13,376
Test	13,376
Mean LoS	3.69
Median LoS	2.27
Number of features	35
Temporal	17
Demographic (static)	18
Number of hospitals (clients)	189

B. Model architecture

For both the centralized and federated training tasks in this work, the Gated Recurrent Unit (GRU) [5], and Long Short-Term Memory (LSTM) [10] networks are used. Both GRU and LSTM belong to the class of Recurrent Neural Networks (RNN) which are widely used when dealing with sequential data. The GRU architecture is comprised of two sole gates. The reset and update gates, respectively denoted as r_t and z_t in (1), resulting in reduced computational complexity. This, in turn, is a desirable characteristic for FL applications where local computational resources and communication overhead pose real challenges [12].

In (1), the governing equations for the GRU cell are shown. These outline the computations that occur for every discrete time step t in the input sequence.

$$\begin{aligned} r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ \tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \tag{1}$$

Similarly, in (2), the governing equations for the LSTM cell are shown. LSTM is comprised of three gates rather than the two gates present in GRU which are denoted as f_t, i_t and o_t, representing the forget gate, input gate and output gate.

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
h_t &= o_t * \tanh(C_t)
\end{aligned} \tag{2}
$$

In both (1) and (2) σ represents a sigmoid neural network layer whereas $\tanh$ represents a $\tanh$ neural network layer with the exception of the last step for LSTM where h_t is calculated. Here, $\tanh$ corresponds to the $\tanh$ activation function such that the values are transformed to fall in $[-1, 1]$. A visual representation of the GRU and LSTM cells is shown in Fig. 3.

For both architectures, x_t represents the input at time step t, which corresponds to one hour in this work. In addition, h_t, h_{t-1} denote the hidden states at time $t, t-1$ respectively and $*$ represents the Hadamard product.

The hidden state h_t, i.e., the output of the cell, is provided as the input of a nonlinear Fully Connected Network (FCN), which yields a single output value representing the predicted LoS. The nonlinearity stems from the ReLU activation function leveraged in the FCN as shown in (3).

$$
\hat{y}_t = ReLU\left(W_{y_t} h_t + b_{y_t}\right), \tag{3}
$$

with $\hat{y}_t$ the predicted value for LoS at time t. Employing $ReLU(x) = max(0, x)$ forces the outcome to be strictly positive. It is impossible for a patient to have a negative LoS, therefore we restrict the model to only yield predictions in the positive domain.

C. Client Recruitment

Consistent with the work discussed in [24], we consider client recruitment to be a mechanism which is to be invoked prior to establishing a federation for training. To this extent, consider the set S of c potential clients with each a local dataset $D_c = \{(x_i, y_i)\}_i$ where x_i denotes the input data and y_i the target.

To facilitate client recruitment, we let each potential client in S report a tuple (P_{co}, n_c) to the central server where P_{co} denotes the local distribution of the target and n_c the local sample size. From here, the global sample size n_g and output distribution P_{go} can be calculated as shown in (4).

$$
n_g = \sum_c n_c, \quad P_{go} = \sum_c P_{co} \tag{4}
$$

Using the tuple reported by the potential client, the local representativeness of the client data ν_c in relation to the global dataset is calculated as a function of the output distribution divergence and the local sample size:

$$
\nu_c = \gamma_{dv} \underbrace{\left| \frac{P_{go}}{n_g} - \frac{P_{co}}{n_c} \right|}_{\beta} + \gamma_{sa} n_c^{-0.5}, \tag{5}
$$

where γ_{dv} and γ_{sa} denote weight parameters that influence the importance of the divergence of the output distribution and local sample size respectively. Furthermore, $\tilde{P}_c - P_c$ converges to $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ at the rate of $O(n_c^{-0.5})$, with $\tilde{P}_c$ the empirical local distribution [11], [26]. From this follows that as n_c grows larger, $\tilde{P}_c$ better approximates P_c [24]. Which is a desired characteristic for client recruitment. By inclusion of the term $n_c^{-0.5}$ in (5), clients with larger local sample sizes are favored over those with smaller sample sizes.

The local distribution divergence in terms of the target denoted as β in (5) is calculated as the difference between the normalized class counts locally and globally. In this work, the target corresponds to the patient LoS in fractional days. To facilitate the computation of the distribution divergence, ten bins are constructed with each bin corresponding to the frequency of target values (LoS in fractional days) within the range for a given bin. The bins are defined as: $[(0, 1), [1, 2), [2, 3), ..., [7, 8), [8, 14), [14, +\infty)]$. This formulation converts the target from continuous to discrete classes for which the class counts are used to compute β in (5).

To select clients for recruitment, the per client representativeness values from (5) are sorted and stored in the vector $\boldsymbol{\nu}$. Consider:

$$
\nu_g = \sum_c \nu_c, \tag{6}
$$

where ν_g represents the global representativeness, used to define the recruitment threshold $\iota = \gamma_{th} \nu_g$ with γ_{th} a configurable hyperparameter. Next, by summing over $\boldsymbol{\nu}$, the value ν_c at which the threshold ι is crossed, is identified. All the corresponding clients for values up until that point in $\boldsymbol{\nu}$ are recruited for the federation. This yields a subset of clients which are the most representative in terms of the target distribution divergence and local sample size.

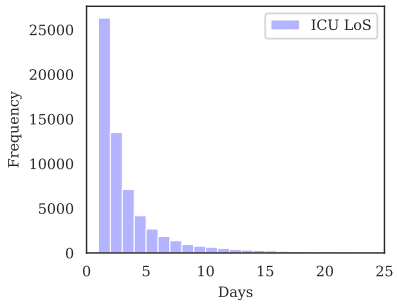

Fig. 1. The distribution of the target (LoS) in the global training data with a cutoff at 25 days on the x-axis. The y-axis represents the number of patients (frequency) with their corresponding LoS in days as indicated on the x-axis.

Intuitively, when applied to the LoS setting, the central server obtains a tuple (P_{co}, n_c) from each of the potential clients to compute n_g and P_{go} which is visually represented as the global target distribution in Fig. 1. This information is then used by the central server to compute for each of the potential clients the local representativeness ν_c as a weighted function in terms of divergence from the global target distribution and

local sample size as shown in (5). A small subset of the local target distributions is shown in Fig. 2. Visually it is clear how some of the local target distributions diverge to a greater extent from the global target distribution, shown in 1, compared to others. In addition, client recruitment requires only a single calculation for each of the clients prior to forming the federation and therefore does not consume any further resources during the remainder of the training procedure.

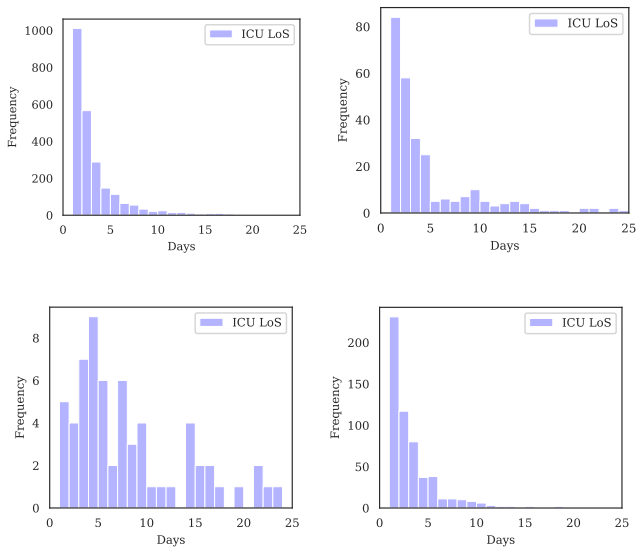

Fig. 2. Subset of the, in total 189, local target distributions in terms of Length of Stay in days and the corresponding patient counts. The distributions shown correspond to, from left to right, the hospitals with the identifiers; 24, 74, 100 and 143.

D. Experimental settings

At training time, all training procedures use *AdamW* [14] for optimization and the Mean Squared Logarithmic Error (MSLE) as the loss function, calculated as:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i+1) - \log(\hat{y}_i+1))^2, \qquad (7)$$

with y_i the true target value and $\hat{y}_i$ the predicted value. Furthermore, the model hyperparameters are fixed over all training iterations, both central and federated. The exact settings are shown in Table II with with L the number of layers, N the hidden dimension for each of the layers, η the learning rate, m the batch size, wd the weight decay for the *AdamW* optimizer and r the dropout.

TABLE II
MODEL HYPERPARAMETER SETTINGS USED FOR BOTH CENTRAL AND FEDERATED TRAINING.

Model	L	N	η	m	wd	r
GRU	2	32	0.005	128	0.005	0.05
LSTM	2	32	0.005	128	0.005	0.05

For evaluation of the performance, all resulting models are evaluated against the hold-out test set containing data from all 189 hospitals. In addition to the MSLE, models are evaluated using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and the Mean Squared Error (MSE), shown in (8). As an indication of the time complexity, the training time is reported, denoted as τ, in seconds.

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \\ MAPE &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \\ MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned} \qquad (8)$$

1) Central training: Central training is performed consistent with the traditional DL procedure in which all data is assumed to be centrally available. The architecture presented in IV-B is trained for a predetermined amount of 15 epochs using the global train and validation sets, which consist of the accumulated data over all potential clients, i.e., the 189 originating hospitals. The resulting central model is subsequently evaluated against the hold-out test set.

2) Federated training: The FL procedure with and without client recruitment is simulated as a single process using the *FedML* framework as proposed in [9]. In this work, for FL without client recruitment, clients are either all considered in each round of training or a subset is randomly sampled consistent with the standard client selection implementation described in *FedAvg* [15]. For FL with client recruitment, the client recruitment process described in IV-C is invoked prior to initiating the federation. Following the described implementation, the recruitment of clients is influenced by three user-defined hyperparameters γ_{dv}, γ_{sa} and γ_{th} which respectively define the importance of the divergence in the target distribution, the local sample size and percentage of the global representativeness to be covered by the recruited clients. In Fig. 3, a general overview of the federated training procedure with client recruitment is shown with M representing the central server that initially recruits clients for the federation after which the federated training procedure is started for training either the GRU or LSTM network.

Four different federated strategies are implemented and trained with either the GRU or LSTM model. The four approaches differ in terms of the number of clients that partake in the federation, denoted as ϵ, the percentage of clients in the federation that contribute to each training round, denoted as δ, and whether the federation is comprised of recruited clients, or all clients. For each of the federated models, each client trains for four epochs per round of server-client communication for a total of 15 rounds. The resulting model is subsequently evaluated against the hold-out test set.

The specifications for the four different FL strategies are; (i) **Fed-AC:** all clients make up the federation and partake in each training round, (ii) **Fed-SC:** all clients make up the

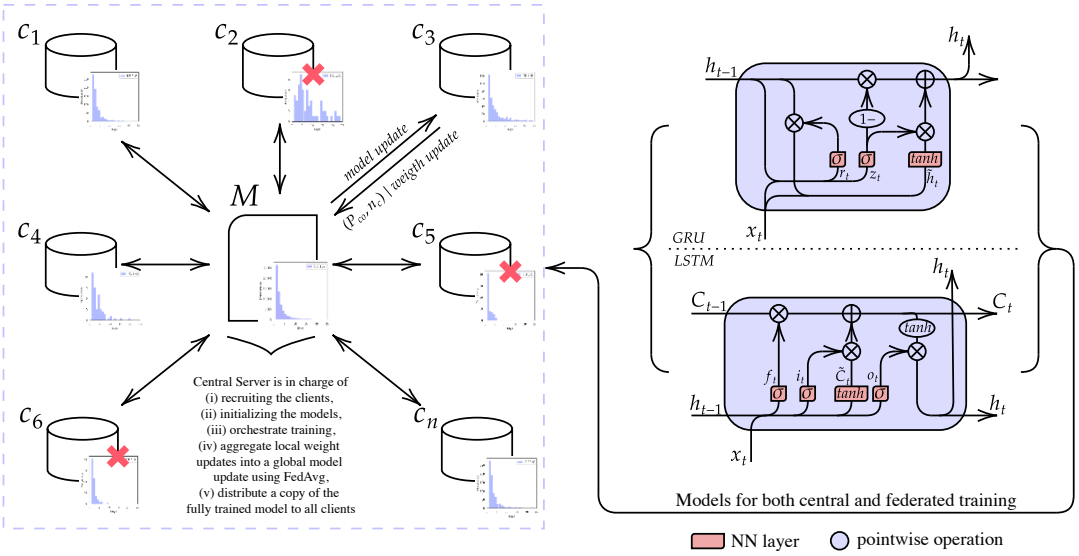

Fig. 3. General overview of the client recruitment and federated training procedure with M the central server and $c_1, ..., c_n$ the potential clients. Each of the clients is depicted with their corresponding target distributions for which the red cross indicates clients that are discarded to participate in the federation based on the recruitment procedure. M is initially in charge of recruiting clients based on the representativeness of the locally hosted data and will subsequently orchestrate the federated training procedure for either the GRU or LSTM models. The GRU and LSTM cells included are a visual representation of the governing equations in (1) and (2).

TABLE III
SETTINGS SPECIFIC TO THE FEDERATED STRATEGIES WITH ϵ THE TOTAL NUMBER OF CLIENTS IN THE FEDERATION, δ THE NUMBER OF CLIENTS RANDOMLY SAMPLED FROM ϵ IN EACH ROUND OF TRAINING AND (γ_{dv}, γ_{sa}, γ_{th}) THE HYPERPARAMETERS FOR CLIENT RECRUITMENT

Strategy	ϵ	δ	γ_{dv}	γ_{sa}	γ_{th}
Fed-AC	189	189	-	-	-
Fed-SC	189	19	-	-	-
Fed-ARC	54	54	0.5	0.5	0.1
Fed-SRC	54	5	0.5	0.5	0.1

federation, 10% of which are randomly sampled to partake in each training round, (iii) **Fed-ARC:** recruited clients make up the federation and partake in each training round and (iv) **Fed-SRC:** recruited clients make up the federation, 10% of which are randomly sampled to partake in each training round. The settings specific to each of the strategies in the federated setting are summarized in Table III.

V. RESULTS AND DISCUSSION

The results obtained for both the central and federated training procedures as described in IV-D1 and IV-D2 are reported in Table IV with τ the training time in seconds and the best performance per metric listed in bold.

A. Client recruitment results

When considering the results reported in Table IV, the best performance for MSLE amongst the federated models is obtained by Fed-AC for both GRU and LSTM. Here, all clients make up the federation and partake in each round of training. However, considering all clients at each round of training incurs significant overhead, as illustrated by the high training time for Fed-AC. The second best performing model in terms of MSLE is Fed-SRC, trained with recruited clients, for which a subset was randomly sampled at each round of training. More specifically, as seen in Table III, 54 clients were recruited from which 5 were randomly sampled at each round of training. For both GRU and LSTM, Fed-SRC outperforms Fed-SC and Fed-ARC in terms of MAE, MSE, MSLE and training time. For GRU, Fed-SRC is the best overall performer in terms of MAE and training time, whereas the best MAE for LSTM is obtained by Fed-AC at 2.21 days, nevertheless, Fed-SRC yields an MAE of 2.22 days at a fraction of the training time.

The improved training time is a direct result of the reduced number of clients recruited for the federation. The improved performance can be attributed to the fact that model training is subject to less noise in the data, again as a direct result of the client recruitment procedure. Therefore, more informative model updates are produced at each round of training, resulting in lower empirical loss, which in turn improves performance.

In summary, the client recruitment approach allows for models to be trained that outperform the standard FL approach (Fed-SC) and perform on par or better than the centrally trained model depending on the metric of interest. In addition, the total training time is drastically reduced compared to either central or federated models without client recruitment. In this use case, models with a lower MAE in combination with lower training time are of most value to the ICU. For both GRU and LSTM, Fed-SRC outperforms the standard Fed-SC

TABLE IV

MODEL PERFORMANCE FOR CENTRAL AND FEDERATED MODELS WITH AND WITHOUT CLIENT RECRUITMENT. STATISTICAL SIGNIFICANCE AMONG THE FEDERATED MODELS IN COMPARISON TO FED-SC IS INDICATED AS * AT THE 5% SIGNIFICANCE LEVEL AND ** AT THE 1% SIGNIFICANCE LEVEL.

Model	Strategy	MAE	MAPE	MSE	MSLE	τ(s)
GRU	Central	2.21 ± 0.02	0.57 ± 0.06	21.94 ± 0.63	0.33 ± 0.01	2128 ± 18
	Fed-AC	2.26 ± 0.06	0.63 ± 0.08**	**21.61 ± 0.73****	**0.33 ± 0.02****	5231 ± 29**
	Fed-SC	2.26 ± 0.06	0.46 ± 0.06	23.98 ± 1.26	0.41 ± 0.05	1469 ± 35
	Fed-ARC	2.27 ± 0.12	0.57 ± 0.17*	22.67 ± 1.83**	0.37 ± 0.05*	3359 ± 25**
	Fed-SRC	**2.21 ± 0.03****	**0.46 ± 0.03**	23.49 ± 0.73	0.37 ± 0.03*	**965 ± 24****
LSTM	Central	2.20 ± 0.01	0.52 ± 0.04	22.45 ± 0.45	0.34 ± 0.01	1891 ± 12
	Fed-AC	**2.21 ± 0.04****	0.48 ± 0.06	**23.18 ± 0.52****	**0.37 ± 0.02****	4663 ± 46**
	Fed-SC	2.28 ± 0.08	0.44 ± 0.03	24.36 ± 1.13	0.43 ± 0.06	1301 ± 20
	Fed-ARC	2.23 ± 0.02	**0.43 ± 0.01**	24.05 ± 0.48	0.40 ± 0.02	2866 ± 24**
	Fed-SRC	2.22 ± 0.03**	0.44 ± 0.02	23.96 ± 0.85	0.39 ± 0.03*	**864 ± 13****

approach, obtaining a better MAE in a fraction of the required training time. Thus, illustrating practical relevance of the client recruitment procedure for larger federations in a real-world, privacy-sensitive, setting.

B. Recruitment parameter effects

To gain additional insight in the behaviour of the client recruitment procedure proposed in IV-C, we assess performance under different settings for the user defined hyperparameters γ_{dv} and γ_{sa}. γ_{dv} influences the importance of the divergence in the distribution between the local target distribution and that of the target in the global data whereas γ_{sa} affects the importance of the local sample size in the client recruitment approach.

We perform this analysis using the GRU model and tune the parameters such that distribution divergence is prioritized over the local sample size and vice versa as follows: **Fed-SRC-QG**: divergence over sample size, and **Fed-SRC-DG**: sample size over divergence. The parameter settings for both the approaches are listed in Table V.

TABLE V

PARAMETER SETTINGS FOR THE QUALITY GREEDY (QG) AND QUANTITY GREEDY (DG) APPROACHES.

Strategy	γ_{dv}	γ_{sa}
Fed-SRC-QG	1	0.01
Fed-SRC-DG	0.01	1

The above formulation is intuitively equivalent to the recruitment process being quality greedy (QG) in the former and quantity (data) greedy (DG) in the latter. The quality greedy recruitment strategy allows for clients with smaller sample sizes for which the output does not diverge significantly in distribution to be recruited. The contrary is true for the quantity greedy strategy in which the recruitment process neglects, to some extent, the distribution divergence in the output and favorably ranks clients with large local samples. Essentially, this approach explores performance in the extremes of the weighted function presented in (5).

The results in Table VI show that neither the quality greedy approach, nor the quantity greedy approach perform better than the Fed-SRC model shown in Table IV. In addition,

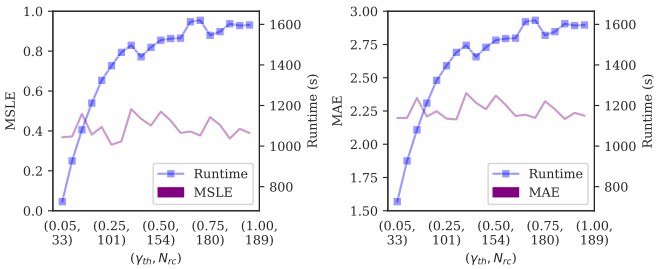

Fig. 4. Runtime versus MSLE (left) and runtime versus MAE (right) in function of gradually increasing values for γ_{th}. Corresponding to a gradually increasing number of recruited clients, denoted as N_{rc}, for each of the training iterations.

we note how the data greedy approach results in increased training time due to the larger sample sizes in the local data of the recruited clients. The findings reported in Table VI, when compared to the results for Fed-SRC in Table IV, show the value of combining both divergence and local sample size in the client recruitment process as neither of the extreme strategies outperform the combined approach.

TABLE VI

MODEL PERFORMANCE FOR FEDERATED TRAINING WITH DATA GREEDY AND QUALITY GREEDY RECRUITMENT STRATEGIES.

Strategy	MAE	MSLE	τ(s)
Fed-SRC-QG	2.23 ± 0.03	0.40 ± 0.03	891 ± 25
Fed-SRC-DG	2.22 ± 0.06	0.39 ± 0.05	1137 ± 15

Having covered the effects on performance under different settings for γ_{dv} and γ_{sa}, we investigate how different settings for γ_{th} affect the performance and training time. Higher values for γ_{th} directly correspond to more recruited clients for the federation, which in turn corresponds to higher training time. To this extent, the value for γ_{th} is gradually increased in steps of 0.05. For each step we observe the performance on the test set in terms of MSLE, MAE and training time as shown in Fig. 4. This shows how there is no direct relation between performance and a higher number of recruited clients, even more so, close to optimal performance can be obtained for low values of γ_{th}, i.e., with few of the most representative clients making up the federation.

VI. Conclusion, Limitations & Future Work

A. Conclusion

In this work, we present a client recruitment approach considering only the local output distribution and local sample size. In addition, we show practical relevance of the proposed method in the medical setting. By recruiting clients in function of the herein defined client-level representativeness, those clients with smaller sample size in combination with those for which the output distribution vastly diverges compared to that of the global data are pre-excluded from the set of potential clients for the federation. By applying client recruitment, the predictive performance of the federated models significantly increases compared to the models trained with the standard FL approach. In addition, training time was significantly reduced as a direct result of the reduced number of clients that partook in training.

B. Limitations & Future Work

The main limitation of this work stems from the introduction of the recruitment parameter, γ_{th}, which directly affects the number of clients recruited for the federation. In a real-world setting, tuning this parameter is not always feasible. In addition, the present work is executed in a simulated, single process environment for which communication overhead is not a factor.

Future work will evaluate performance in a real-world setting were data is hosted on actual servers corresponding to the hospitals in separate networks. Here, client recruitment is of even greater importance as it can greatly reduce the overall required server-client communication. In addition, a direction for future research is to look at how to, a priori, approximate the optimal setting for γ_{th}. Furthermore, future work will explore alternative recruitment strategies with a focus on sampling from diverse subgroups in the data and assess local performance of the federated models against models trained on the local data only.

References

[1] Abdulrahman Al-Dailami, Hulin Kuang, and Jianxin Wang. Predicting length of stay in ICU and mortality with temporal dilated separable convolution and context-aware feature fusion. *Comput. Biol. Med.*, 151(Pt A):106278, December 2022.

[2] Rohan Bhardwaj, Ankita R. Nambiar, and Debojyoti Dutta. A study of machine learning in healthcare. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 236–241, 2017.

[3] Alison Callahan and Nigam H. Shah. Chapter 19 - machine learning in healthcare. In Aziz Sheikh, Kathrin M. Cresswell, Adam Wright, and David W. Bates, editors, *Key Advances in Clinical Informatics*, pages 279–291. Academic Press, 2017.

[4] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/, 1996.

[5] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[6] Mario Ciampi, Mario Sicuranza, and Stefano Silvestri. A privacy-preserving and standard-based architecture for secondary use of clinical data. *Information*, 13(2), 2022.

[7] European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016.

[8] A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, June 2000.

[9] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. Fedml: A research library and benchmark for federated machine learning. *Advances in Neural Information Processing Systems, Best Paper Award at Federate Learning Workshop*, 2020.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[11] Tito Homem-de Mello. On rates of convergence for stochastic optimization problems under non–independent and identically distributed sampling. *SIAM Journal on Optimization*, 19(2):524–551, 2008.

[12] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *CoRR*, abs/1908.07873, 2019.

[13] Wenqi Li, Fausto Milletari, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M. Jorge Cardoso, and Andrew Feng. Privacy-preserving federated brain tumour segmentation. *CoRR*, abs/1910.00962, 2019.

[14] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[16] Arash Mehrjou, Ashkan Soleymani, Annika Buchholz, Jürgen Hetzel, Patrick Schwab, and Stefan Bauer. Federated learning in multi-center critical care research: A systematic case study using the eicu database. *CoRR*, abs/2204.09328, 2022.

[17] Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–7, 2019.

[18] Giovanni Paragliola. A federated learning-based approach to recognize subjects at a high risk of hypertension in a non-stationary scenario. *Information Sciences*, 622:16–33, 2023.

[19] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, Sep 2018.

[20] Tom Joseph Pollard, Alistair Edward William Johnson, Jesse Raffa, and Omar Badawi. The eICU collaborative research database, 2017.

[21] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119, Sep 2020.

[22] Emma Rocheteau, Pietro Liò, and Stephanie Hyland. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 58–68, New York, NY, USA, 2021. Association for Computing Machinery.

[23] Emma Rocheteau, Catherine Tong, Petar Velickovic, Nicholas D. Lane, and Pietro Liò. Predicting patient outcomes with graph representation learning. *CoRR*, abs/2101.03940, 2021.

[24] Yichen Ruan, Xiaoxi Zhang, and Carlee Joe-Wong. How valuable is your data? optimizing client recruitment in federated learning. In *2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pages 1–8, 2021.

[25] K. Shailaja, B. Seetharamulu, and M. A. Jabbar. Machine learning in healthcare: A review. In *2018 Second International Conference on*

Electronics, Communication and Aerospace Technology (ICECA), pages 910–914, 2018.

[26] Alexander Shapiro. Monte carlo sampling methods. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 353–425. Elsevier, 2003.

[27] Micah J. Sheller, G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 92–104, Cham, 2019. Springer International Publishing.

[28] Arnon Vandenberghe, Lyse Naomi Wamba Momo, Vincent Scheltjens, and Bart De Moor. Multimodal deep learning for early length of stay prediction using patient similarity embeddings. In *Proc. of BNAIC/BeNeLearn*. Mechelen, Belgium, 2022.

[29] Wenchao Xia, Tony Q. S. Quek, Kun Guo, Wanli Wen, Howard H. Yang, and Hongbo Zhu. Multi-armed bandit-based client scheduling for federated learning. *IEEE Transactions on Wireless Communications*, 19(11):7108–7123, 2020.

[30] Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2):1188–1200, 2021.

[31] Naoya Yoshida, Takayuki Nishio, Masahiro Morikura, and Koji Yamamoto. Mab-based client selection for federated learning with uncertain resources in mobile networks. In *2020 IEEE Globecom Workshops (GC Wkshps*, pages 1–6, 2020.

[32] Hangjia Zhang, Zhijun Xie, Roozbeh Zarei, Tao Wu, and Kewei Chen. Adaptive client selection in resource constrained federated learning systems: A deep reinforcement learning approach. *IEEE Access*, 9:98423–98432, 2021.

[33] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *CoRR*, abs/2106.06843, 2021.

APPENDIX

A. Features

Table VII outlines and describes the temporal and static features that constitute the data cohort used for all training procedures described within this work.

TABLE VII
EXTRACTED TEMPORAL AND STATIC FEATURES

Type	Feature	Description
Temporal	FiO2	Patient's FiO2 value
	Bedside glucose	Patient's glucose level
	Cvp	Patient's cvp value
	Heartrate	Patient's heart rate value
	Noninvasivediastolic	Patient's non invasive diastolic value
	Noninvasivemean	Patient's non invasive mean value
	Noninvasivesystolic	Patient's non invasive systolic value
	Respiration	Patient's respiration value
	Sao2	Patient's spO2 value
	St1	Patient's st1 value
	St2	Patient's st2 value
	St3	Patient's st3 value
	Systemicdiastolic	Patient's diastolic value
	Systemicmean	Patient's mean pressure
	Systemicsystolic	Patient's systolic value
	Temperature	Patient's temperature value in celsius
	Hour	Time since admission
Static	Hospitalid	Surrogate key for the hospital
	Gender	Gender of the patient
	Age	Patient's age in full years
	Admissionheight	Admission height of the patient in cm
	Admissionweight	Admission weight of the patient in kg
	Intubated	Whether patient is intubated at the time of the worst ABG result
	Vent	Whether patient is ventilated at the worst respiratory rate
	Dialysis	Whether patient is on dialysis
	Eyes	GCS score (1 to 4)
	Motor	GCS score (1 to 5)
	Verbal	GCS score (1 to 6)
	Meds	Whether GCS score could not be obtained due to meds
	Ethnicity	Patient's ethnicity
	Unittype	The picklist unit type of the unit
	Unitadmitsource	Picklist location from where the patient was admitted
	Unitstaytype	Patient's unit stay type
	Physicianspeciality	Picklist specialty of the care provider
	> 89	Whether patient is over 89 years old