Power-Law Convergence in Federated Learning for Distributed Residential Load Forecasting

Alexander Wallis*, Sascha Hauke†, Hannah Jörg**, Konstantin Ziegler*
*Department of Computer Science
**Department of Interdisciplinary Studies
University of Applied Sciences Landshut
Landshut, Germany
e-mail: {alexander.wallis | hannah.joerg | konstantin.ziegler}@haw-landshut.de
†Department of Computer Sciences
Justus Liebig University Giessen
Giessen, Germany
e-mail: sascha.hauke@uni-giessen.de

***Abstract*—The integration of renewable energy resources transforms traditional energy systems, introducing prosumers entities that both produce and consume energy as key participants in modern Smart Grids. Effective load forecasting is mandatory for optimizing energy resources and grid stability. Federated Learning has emerged as a promising approach for distributed training of Machine Learning-based forecasting models. This enables collaborative model optimization across multiple prosumers while preserving data privacy. However, the impact of unbalanced data sets across participants remains a critical challenge in terms of potentially affecting learning convergence and forecast accuracy. In this work, we define and implement a Federated Learning system based on real-world electricity consumption data from a variety of prosumers. Experimental results demonstrate the trade-off between centralized and federated learning approaches, providing insights into addressing data heterogeneity in Federated Learning systems. Additionally, we show that the models convergence during training with unbalanced data sets follows a power law function. These insights highlight the potential of Federated Learning to support the evolution of distributed energy systems while ensuring data-privacy and scalability. Furthermore, the results provide actionable insights for grid operators balancing privacy, efficiency, and accuracy. Future research directions include other strategies to mitigate the effect of data imbalances and further improve the efficiency of federated optimization for dynamic energy systems.**

Keywords*-*Short-Term Load Forecasting; Federated Learning; Smart Grid; Data Privacy; Distributed Data.

I. Introduction

This work extends the results of our conference paper [1], which provides a distributed approach for Short-Term Load Forecasting (STLF) on residential household level with respect to data privacy. Accurate load forecasting is mandatory for stable and reliable Smart Grid (SG) operation. But, the accuracy of load forecasting models, in particular Machine Learning (ML) based models, highly depends on the amount and quality of available training data [2]. Especially on smaller grid levels, e.g., low-voltage grids, or even residential household levels, the available electricity consumption data are very limited. But, with the rise of *prosumers* – consumers also able to produce electricity – prediction models on exactly this grid level are crucial for network management tasks [3].

Even if households are able to record and transmit electricity consumption data through smart meter utilization, the grid operator needs sufficient data storage and computational resources to process the data. Otherwise, the gathered data must be transferred for further processing. This transfer raises data privacy concerns and is even prohibited by law, e.g., General Data Protection Regulation [4]. The ability of information and behavior retrieval based on leakage of electricity consumption data has already been shown in the past [5], [6], [7].

Here, Federated Learning (FL) seems to be a promising approach to develop a single ML model for electricity consumption forecasting with distributed data sets – and at the same time satisfying data privacy regulation [8]. In contrast to the traditional approach, where the training of the ML model is done centralized, this task is shifted to each user individually.

In [9], FL was first used by McMahan et al. to train prediction models on mobile devices through users' keyboard inputs. Afterwards, applications with FL were proposed in various fields, e.g., medical and health care, industrial engineering, finance, transportation [10], [11], [12].

For SG development, various FL approaches were proposed, too. In [13], FL is used for anomaly detection in terms of energy usage with a detection rate compared to centralized approaches. The authors in [14] present a conceptual framework for secure FL usage in SG environments with focus on vertical and horizontal data distribution over the clients. A detailed overview of further interesting FL researches in the field of SGs is given in [15].

Although FL can be a promising approach for distributed load forecasting, the impact of unbalanced data sets among the clients is unclear. To evaluate FL in the context of prosumer-level load forecasting, we present the following contributions in this work:

- Definition and implementation of FL system composed of a variety of prosumer based on real-world electricity consumption data.

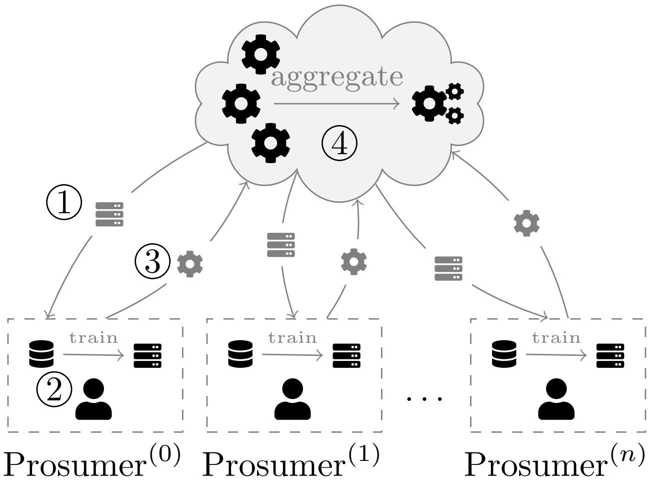


Figure 1. In a Federated Learning approach, all prosumers train their models locally on their own data.

- Comparison of forecast accuracy between a centralized and a federated learning approach for model optimization.
- Investigation of the influence of unbalanced data sets within a federation on the learning convergence and the overall forecasting error.
- Analysis of the relation between number of unbalanced clients and training convergence.

This work is organized as following. First, the necessary background information as well as notation and terminologies are given in Section II. Second, the proposed FL approach is described in detail and the different experiments conducted are described in Section III. Third, the experiment results are presented, compared, and subsequently evaluated and discussed w.r.t. forecasting accuracy in Section IV. Fourth, limitations of the proposed work and solutions are presented in Section V. Fifth and last, the insights gained from the experiments' results are summarized and starting points for further research are given in Section VI.

II. Background

Before further detailing the conducted experiments in Section III, we give the respective problem formulation (Section II-A) and background information on FL (Section II-B) as well as an overview of related work (Section II-C).

A. Problem Formulation

Basically, the load forecasting problem can be categorized into three groups based on the forecast horizon: (i) short-term, (ii) middle-term and (iii) long-term load forecasting. In this work, attention is paid on STLF, since we are interested in a household's next-day electricity consumption.

Traditionally, STLF has been addressed using both statistical and ML techniques. Early approaches include time series models such as Autoregressive Integrated Moving Average (ARIMA) and its variants [16]. With the increasing availability of high-resolution smart meter data, ML methods have gained more focus. Here, the more recent advances rely on neural networks with deep learning architectures [17]. In particular, Long-Short Term Memory Neural Network (LSTM) networks and Gated Recurrent Units (GRU) are widely used for their ability to capture temporal dependencies, whereas Convolutional Neural Networks (CNNs) and hybrid CNN-LSTM models leverage spatial and temporal features [18], [19]. In the following, the fundamental problem formulation for STLF is given.

Let $\mathbf{x}_d = (x_d^{(0)}, ..., x_d^{(T)}) \in R^T$ be a household's consumption of day d divided into T time intervals. Further, let $\mathbf{y}_d = (y_{d+1}^{(0)}, ..., y_{d+1}^{(T)}) \in R^T$ be the next day's electricity consumption, then $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 0, ..., D\}$ is the data set composed of input-output pairs for a total of D days. Now, a supervised learning approach approximates a function $\mathbf{y}_d \approx \hat{f}(\mathbf{x}_d)$ for the following optimization problem:

$$\arg\min_{\hat{f} \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(\hat{f}(\mathbf{x_i}), \mathbf{y}_i) \qquad (1)$$

where $L(\cdot)$ is the desired cost function to be minimized.

Typically, in a centralized learning setting, this is done by collecting each household's data and subsequently by training a combined forecasting model, which is afterwards distributed to every household. Indeed, this rises all of the problems and concerns described earlier (see Section I) and FL is a promising approach to tackle all of them.

B. Federated Learning

Contrary to the centralized learning, a FL approach guarantees data-privacy by preserving prosumers' consumption data locally. A collaboration of prosumer – a so-called *federation* – trains a STLF model by only exchanging respective model parameters. Typically, the participants within a federation are called clients but in this work the terms clients, prosumers and households are used interchangeably. Let $\mathcal{P} = \{p^{(i)} | i = 0, ..., N\}$ be the set of N prosumers then FL procedure involves the following steps:

1) **Distribution** of the initial global model to all prosumers which are part of the federation $p \in \mathcal{P}$.
2) **Training** of the global model by adjusting it's parameters based on the local data set of every prosumer.
3) **Returning** the adjusted model parameters to a central unit, e.g., trusted 3rd party, data center, one of the participants.
4) **Aggregation** of all received parameters by a predefined `aggregate`-function and integration into the global model.

This whole procedure, also depicted in Figure 1, is repeated over a defined number of *communication rounds* r. Interestingly, reducing the number C of clients participating in every learning round increases the communication efficiency without loss of prediction accuracy [9]. So, in every round a prosumer subset $\mathcal{P}'_r \subseteq \mathcal{P}$ with $|\mathcal{P}'_r| = C$ is randomly chosen to take part in the training task in step 2.

Beside the number of prosumers involved in training, the used `aggregate`-function offers additional flexibility. In [9], the author introduces `FedSGD` and `FedAvg`, where the later is the common approach for solving the FL problem by calculating the (weighted) average (often mean) per parameter. Other aggregation approaches are, e.g., federated adaptive optimizers (`FedAdam`, `FedAdagrad`, `FedYogi`) [20], momentum-based variance-reduced technique (`FAFED`) [21], heterogeneity focused (`FedProx` [22], `SCAFFOLD` [23]). There are plenty more proposed `aggregate`-methods, and the related questions in terms of, e.g., applicability, optimality, generalization, are major research topics.

At this point, it is worth noting that additional security mechanism are needed to guarantee some desired security level. Although, FL offers a framework for data-privacy in distributed learning, data leakage or reconstruction attacks are still possible [24]. Privacy enhancing techniques applicable for FL settings are, e.g., differential privacy and homomorphic encryption [25].

In the next section, we give an overview of existing FL research with focus on STLF.

C. Related Work

After describing the FL approach in general, we give an overview of existing FL research conducted in the field of residential STLF. Here, we limit the related work explicitly to (i) residential households and (ii) maximum 24-hour forecast horizon.

A comparison between `FedAvg` and `FedSGD` with different forecast horizons (1 h and 24 h) is given in [26]. They showed that their proposed FL model with `FedAvg` reaches higher accuracy than a centralized and a personalized model.

In [27], the authors compare the forecasting accuracy of a FL model on prosumers involved in training and on hold-out prosumers. They choose this approach to evaluate how well the global model fit for non-participating prosumers. Here, the non-participant prosumers fine tune the pre-trained model for 5 epochs locally. They conclude that this fine tuning step improves the forecast accuracy compared to the global model.

In terms of unbalanced client data distribution, Liu et al. proposed the closest approach [28]. Here, clients are divided into 5 groups based on the resolution of their available consumption data ranging from 300 s to 1.800 s.

A hybrid CNN-LSTM model is used in a FL setting in [29]. To handle the consumption heterogeneity, the authors propose a model fine-tuning step after the weight aggregation based on multiple kernel variant of maximum mean discrepancies. Furthermore, all clients are involved in every training and the number of data samples are equal over all clients.

The authors in [30] compare the accuracy of a centralized model with a FL one, a FL plus clustering, and FL plus clustering and subsequently local fine tuning. Here, the last approach reaches the highest accuracy. But, to manage all experiment permutations the evaluations are done with fixed $C = 0.1$.

A personalized FL approach is presented by Rahman et al. [31]. Here, a meta-learning-based strategy is applied such that each client trains their local LSTM with different learning rates. This strategy is developed to address data heterogeneity among the clients. The provided simulations show that their personalized approach reach higher prediction accuracy than traditional LSTM as well as FL approaches.

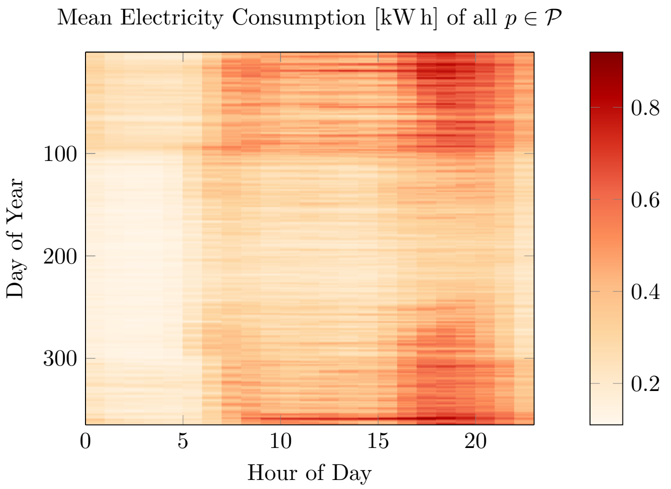


Figure 2. Mean electricity consumption of all selected households from the `SmartMeterInLondon` data set.

All of the mentioned related work are summarized with their respective training and model parameters in Table I. It can be seen that the related work in terms of unbalanced data sets is non existing – as far as we know – for the STLF problem on residential prosumer level.

III. Methodology

To evaluate our proposed FL approach, different experiments are conducted in this work. Therefore, we build a federation composed of prosumers represented by household data taken from public available real-world electricity records (see Section III-A).

A. Used Data Set

In this work, residential household data are taken from the `SmartMetersInLondon` [32] data set, which is a refactored version of the "Low Carbon London Project" data. This data set contains electricity consumption records for 5, 567 London households between November 2011 and February 2014. In the following, the conducted data preprocessing and preparation steps as well as the selection of suitable households is described.

a) Household Selection: Since the date range differs between prosumers in the data set, only houses with the most overlap are selected. Furthermore, households with more than three consecutive hours of missing values are removed – otherwise, missing values are linearly interpolated. In total, 20 households are selected suitable for further usage. The hourly mean electricity consumption is depicted for every day in the training set in Figure 2. Subsequently, the respective consumption data is preprocessed for every selected household in the following.

TABLE I. OVERVIEW AND SUMMARY OF RELATED WORK FOR FEDERATED LEARNING (FL) APPROACHES FOR RESIDENTIAL SHORT-TERM LOAD FORECASTING (STLF).

Related Work	#Clients	C	ML-Model	Data Set	Balanced Data	Aggregation
Taïk and Cherkaoui [27]	200	$5, 10$	LSTM	AUSTIN	yes	FedAVG
Fekri et al. [26]	19	6	LSTM	non-public	yes	FedSDG, FedAVG
Liu et al. [28]	50	10	iQGRU	AUSTIN	semi	FedAVG
Shi and Xu [29]	10	10	CNN-LSTM	LONDON	yes	FedAVG
Briggs et al. [30]	100	0.1	LSTM	LONDON	yes	FedAVG
Rahman et al. [31]	5	5	LSTM	FRANCE	both	FedAVG
our work [1]	20	$1, 2, 5, 10, 20$	MLP	LONDON	no	FedAVG

b) Data Preprocessing: Since the date ranges of available data varies tremendously across all prosumers, we select the time between 1st January 2013 and 28th February 2014 with the most overlapping data. This interval is further divided into train and test data ($\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$), whereas the whole year 2013 is used for training and the remaining data for testing. This leads to $|\mathcal{D}^{(p)}_{\text{train}}| = 8,760$ and $|\mathcal{D}^{(p)}_{\text{test}}| = 1,416$ samples for every prosumer. For every prosumer, both data sets are rescaled individually with the standardization given by

$$x' = \frac{x - \sigma}{\mu}, \tag{2}$$

where x' is the transformed consumption time series with mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

c) Look-back and Forecast Horizon: The accuracy of time series forecasting depends on both, the chosen look-back window as well as the forecast horizon. In the related work (Section II-C), those parameter differ across studies. Here, our proposed forecasting model uses the last 24 h as input to predict the next 24 h. Although, additional features, e.g., weather, holiday, weekday/weekend, can reduce the forecast error, we restrict our model to the raw consumption values. In [33], we evaluate the FL model with further feature engineering. So this leads to an input vector $\mathbf{x} = (x_t, x_{t-1}, ..., x_{t-23}) \in \mathbb{R}^{24}$ and an output vector $\mathbf{y} = (y_{t+1}, y_{t+2}, \dots, y_{t+24}) \in \mathbb{R}^{24}$ for every day and for each prosumer in the data set.

After the household selection and necessary preprocessing steps, the used ML model architecture, as well as further details on the overall development process is given in the next part.

B. System Setting

In this section, we give all the relevant information about the model architecture and used hyperparameters. Afterwards, a definition for different kinds of learning prosumers within the federation based on the ability to store training data is presented. A description of the used federation, as well as the training procedure is given in the third part.

a) Model and Hyperparameters: In this work, we choose a vanilla Multi-Layer Perceptron (MLP) as model architecture, similar to the proposed model in [9]. This architecture allows an easy implementation and training on lightweight devices with limited computational resources. This fully connected MLP has two hidden layers with 200 neurons each and uses a Rectified Linear Unit (ReLU) as activation function.

$$\text{ReLU} = \max(0, x) = \begin{cases} x \text{ if } x > 0 \\ 0 x \leq 0. \end{cases} \tag{3}$$

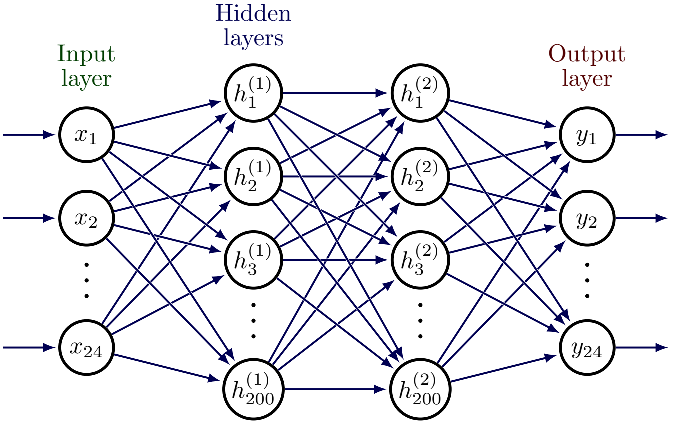


Figure 3. Network architecture used in this work. Fully connected MLP with 2 hidden layers, 200 neurons each and ReLU activation function. The input $\mathbf{x} = (x_{t-24}, \dots, x_t)$.

The final model architecture used in this work for every experiment is illustrated in Figure 3.

b) Weak and Strong Prosumer: We introduce the terms *strong* and *weak* prosumer, to describe two different types of prosumers based on the amount of available training data. The two types are defined the following way:

Definition 1. Let $p \in \mathcal{P}$ be a prosumer only able to store training data between two consecutive communication rounds, then it is called a *weak* prosumer p_{weak}.

Definition 2. Let $p \in \mathcal{P}$ be a prosumer with no storage limitations, then it is called a *strong* prosumer p_{strong}.

Based on the Definitions 1 and 2, we define the fraction of strong prosumers within a federation as the so-called *strong-prosumer-fraction*:

Definition 3. Let $|p_{\text{weak}}|, |p_{\text{strong}}|$ be the number of weak respective strong prosumers in $\mathcal{P}$, then the strong-prosumer-fraction is defined as $\phi = \frac{|p_{\text{strong}}|}{|p_{\text{weak}}| + |p_{\text{strong}}|}$.

This allows a straightforward distinction between prosumers within a federation and introduces another parameter for the overall training procedure.

c) Training Procedure: For all conducted experiments, with or without strong and weak prosumers, the respective training procedure takes $r = 100$ communication rounds in total. At $r = 0$ the global model's weights w are randomly initialized. After every round, the global model's weights are updated by a weighted `FedAvg` aggregation function, given as

$$w_{r+1} \leftarrow \sum_{p \in \mathcal{P}'_r} \frac{n_p}{n} w_r^{(p)}, \tag{4}$$

where n_p, n is the number of sample per prosumer respective the number of all samples. The local weights $w_r^{(p)}$ are calculated locally for every $p \in \mathcal{P}'_r$ in parallel by

$$w_r^{(p)} \leftarrow w_r - \eta \nabla_w \mathcal{L}(w_r; \mathbf{x}_i, \mathbf{y}_i) \tag{5}$$

for a single epoch with a learning rate of $\eta = 0.001$ and the Mean Squared Error (MSE) as loss function $\mathcal{L}(\cdot)$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \tag{6}$$

where n is the number of test set samples and $\hat{y}_i, y_i$ is the predicted respective actual consumption value.

To evaluate the proposed FL approach and also to analyze the impact of unbalanced data sets, various experiments are conducted, which are further detailed in the following section.

C. Experiment Settings

The proposed FL approach for residential STLF is evaluated in different experiments. The evaluation is based on the MSE error metric given in Equation 6. In total, we run the following three experiments:

I **Benchmark** A centralized model – as well as one local model for every prosumer – is trained over r epochs.

II **Number of Learners** Since a new subset of learning prosumers is selected in every round (see Section II-B), we evaluate the model's forecast accuracy for different number of learners $C = \{1, 3, 5, 7, 10, 20\}$.

III **Strong Prosumer Fraction** With the introduction of weak and strong prosumers, we evaluate our FL approach based on unbalanced data sets. For $C = \{1, 10, 20\}$ the strong-prosumer-fraction $\phi = \{0.05, 0.25, 0.5, 0.75, 1\}$ is considered. Here, the unbalanced data set evolves over the communication rounds $r = \{1, 2, \dots, 100\}$ by:

$$\text{weak:} \quad \mathcal{D}_r^{(p)} = \mathcal{D}_{r-1:r}^{(p)} \tag{7}$$

$$\text{strong:} \quad \mathcal{D}_r^{(p)} = \mathcal{D}_{0:r}^{(p)}. \tag{8}$$

So, for strong prosumer the training samples increase by $n = \lfloor \frac{|\mathcal{D}|}{r} \rfloor$ in every round, whereas for weak prosumer the samples have a fixed size of n.

The experiments I-III are repeated for $N = 10$ times to handle the randomness via model initialization and prosumer sampling with C, ϕ. Our proposed FL approach is implemented in `Python=3.9` with `PyTorch` and model training was executed on a local machine with a Nvidia Geforce RTX 2080 graphic card. The experiments' results are listed in the next section.

TABLE II. TEST SET ERROR FOR EXPERIMENT I. MSE IS CALCULATED OVER ALL 20 PROSUMERS

Model	↓ MSE ($\mu \pm \sigma$)	min	max	won
centralized	0.181 ± 0.13	0.030	0.545	3 out of 20
personalized	0.166 ± 0.13	0.021	0.514	17 out of 20

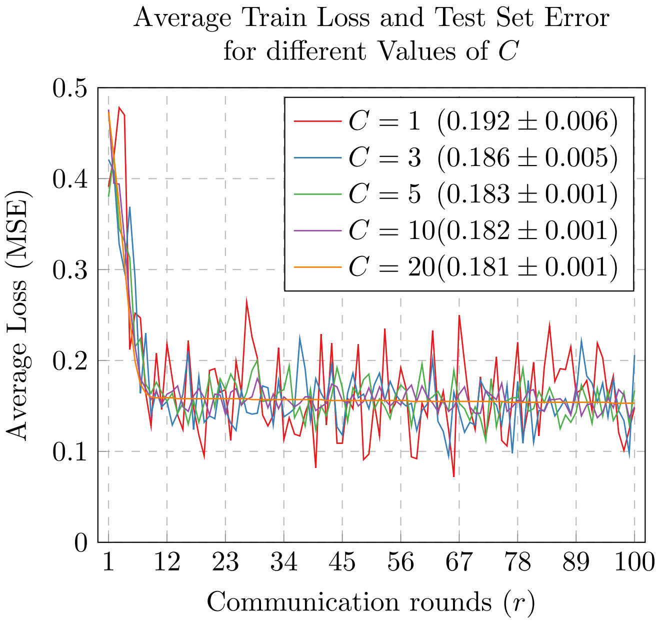


Figure 4. Experiment II: Train loss and test set error with mean and standard deviation over 10 repetitions for different values of C.

IV. Experiment Results & Discussion

The results of the various experiments are presented in the same order as defined in Section III-C. The respective results are provided below, followed by a detailed analysis and discussion.

Figure 4 illustrates the training loss across all communication rounds r as well as the test set error in the legend. For the different values of $C = \{1, 3, 5, 10, 20\}$, the test set error is given as mean with standard deviation over all 10 repetitions. Similar to experiment I, the MSE is calculated over all prosumers $p \in \mathcal{P}$ without individual examination.

For experiment III, results are given in two ways. First, the average training loss over all runs is depicted in Figure 5. Second, Table III lists the test set errors. In addition to numerical values over all prosumers, the MSE is also calculated separately for the sets of p_{weak} and p_{strong}. The minimum and maximum MSE values are determined over all 10 runs combined for each combination of C- and ϕ-values.

In this work, a FL approach was proposed for the STLF problem at residential prosumer level. Three experiments were

TABLE III. TEST SET ERROR FOR EXPERIMENT III. ERROR IS GIVEN AS MSE WITH MEAN AND STANDARD DEVIATION OVER ALL 10 REPETITIONS.

C	ϕ	↓ MSE $(\mu \pm \sigma)$				
		all	strong	weak	min	max
1	0.05	*0.215 ± 0.14*	0.192 ± 0.12	0.216 ± 0.14	0.026	0.674
	0.25	**0.193 ± 0.12**	0.209 ± 0.15	0.188 ± 0.11	0.039	0.597
	0.5	0.194 ± 0.12	0.202 ± 0.13	0.186 ± 0.12	0.037	0.565
	0.75	0.196 ± 0.12	0.194 ± 0.12	0.199 ± 0.13	0.038	0.587
	1	0.201 ± 0.13	0.201 ± 0.13	–	0.036	0.626
10	1	*0.223 ± 0.15*	0.142 ± 0.07	0.227 ± 0.15	0.029	0.750
	0.25	0.186 ± 0.12	0.187 ± 0.13	0.186 ± 0.11	0.033	0.540
	0.5	0.184 ± 0.12	0.182 ± 0.12	0.187 ± 0.12	0.038	0.550
	0.75	0.181 ± 0.12	0.185 ± 0.12	0.170 ± 0.10	0.038	0.525
	1	**0.180 ± 0.12**	0.180 ± 0.12	–	0.041	0.527
20	1	*0.198 ± 0.13*	0.205 ± 0.13	0.198 ± 0.13	0.034	0.711
	0.25	0.190 ± 0.12	0.193 ± 0.13	0.189 ± 0.12	0.035	0.591
	0.5	0.183 ± 0.12	0.173 ± 0.11	0.192 ± 0.13	0.040	0.546
	0.75	0.181 ± 0.12	0.172 ± 0.11	0.208 ± 0.12	0.038	0.523
	1	**0.179 ± 0.11**	0.179 ± 0.11	–	0.042	0.516

Note: lowest error is in **bold**, highest in *italic*.

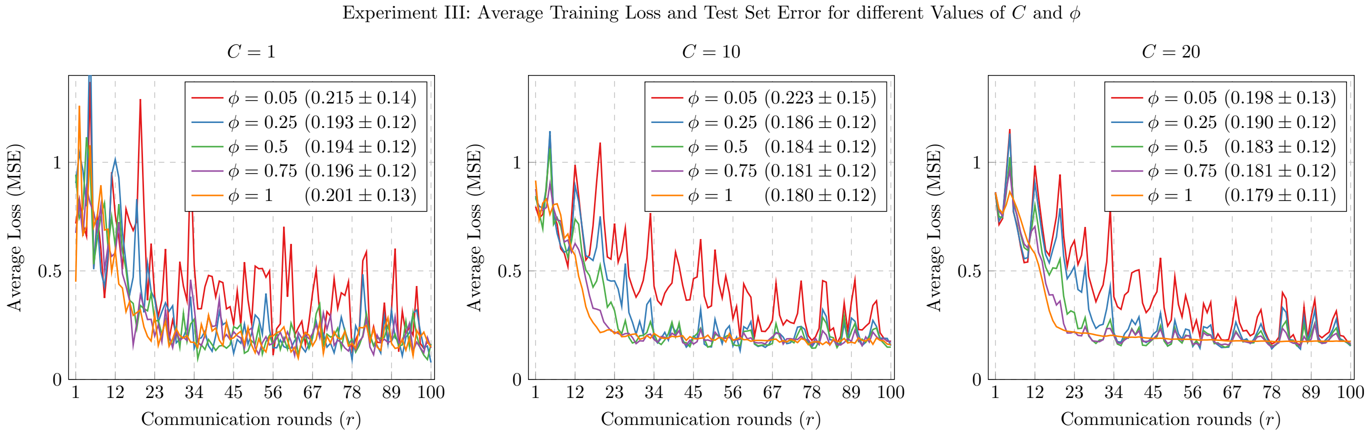


Figure 5. The training loss and test set error for different fractions of strong prosumer ϕ evaluated for $C = 1$ (left), $C = 10$ (middle), and $C = 20$ (right).

conducted to analyze the impact of unbalanced data distribution among prosumers within the federation.

The first experiment compared a centralized MLP trained on all prosumers' data with a personalized MLP trained individually for each prosumer. Of 20 households in total, 17 times the personalized model reaches a higher accuracy (see Table II). This indicates a strong distribution of consumption behaviour across the prosumers since more data does not guarantee better results.

The second experiment examined the effect of different numbers of learners. As shown in Figure 4, test set errors show minimal variation for $C > 1$, with nearly identical training loss reduction. However, lower C-values introduce more variance, emphasizing trade-off between distribution computational resources and learning efficiency.

In real-world scenarios, training data availability varies among prosumers due to recording and storage capabilities as well as temporal offsets in joining the federation. To address this, the third experiment introduced the distinction between weak and strong prosumers, defined by storage capability.

TABLE IV. SUMMARY OF COMMUNICATION ROUNDS TO REACH TARGET MSE FOR 10 RUNS WITH MEAN (μ) AND STANDARD DEVIATION (σ).

C	ϕ	Rounds to target MSE			
		Mean (μ)	STD (σ)	Min	Max
10	0.05	64	14	48	84
	0.25	31	6	24	42
	0.5	24	2	22	27
	0.75	21	3	16	26
	1	18	2	15	22
20	0.05	60	16	48	98
	0.25	34	8	26	55
	0.5	26	2	20	28
	0.75	22	2	20	26
	1	31	2	29	35

The strong prosumer fraction ϕ represents the proportion of strong prosumers within a federation. Figure 5 indicates slower training convergence with a decreasing number of strong prosumers, irrespective of C-values. However, reducing ϕ to 0.75 or 0.5 did not significantly impact training speed or test set error. This finding is relevant for practical applications, suggesting that not all prosumers need to contribute learning resources to maintain overall performance. This is further discussed in the following part.

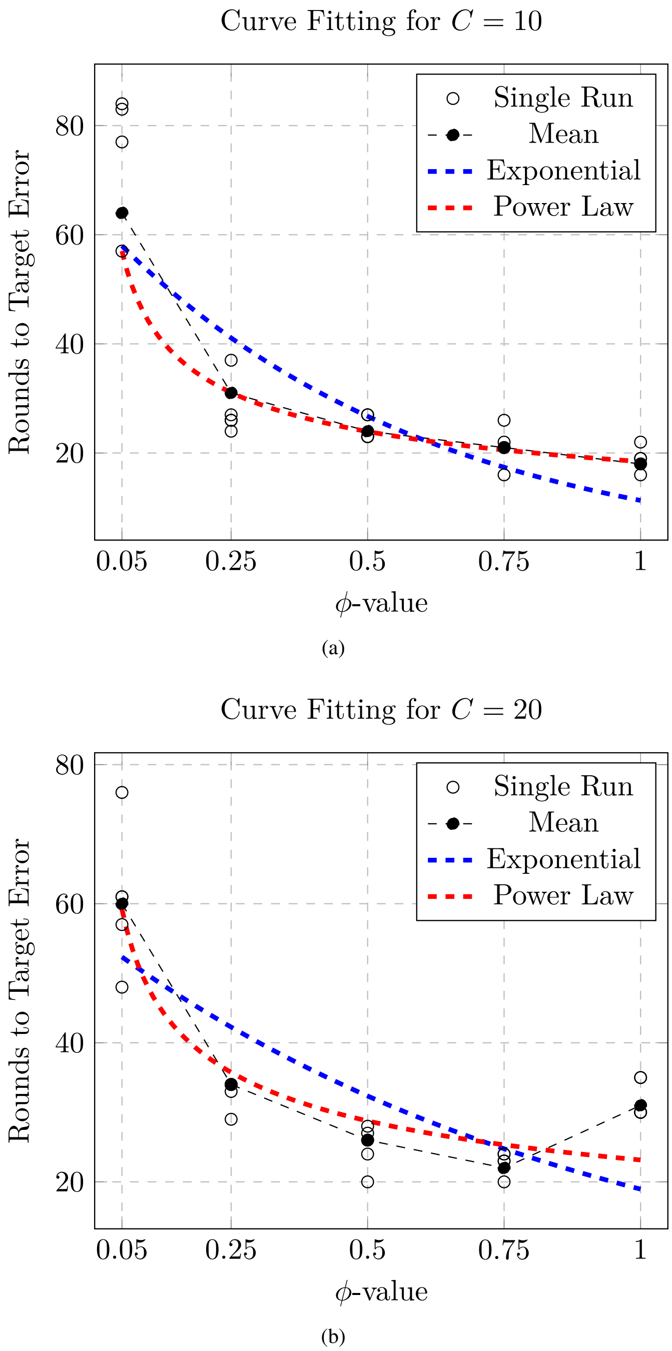
A. Impact of Strong Prosumer Fraction

To evaluate the impact of stron prosumer fraction ϕ on the overall training convergence, we introduce a target error $\text{MSE}_{\text{target}} = 0.20$. This value is chosen based on the test set error from Table III, where all lowest errors are below this threshold. Afterwards, we determine the communication rounds needed to reach the desired train error $\text{MSE}_{\text{train}} < \text{MSE}_{\text{target}}$. This is repeated for all ϕ-values and every single run. The mean (μ) numbers of needed communication rounds are listed in Table IV with respective standard deviations (σ). Where, we only consider $C = 10$ and $C = 20$ since for the $C = 1$ case, the variance over all runs is too high to get meaningful results.

From Table IV, we can see that with increasing ϕ-value the number of communication rounds to reach the target MSE is decreasing. This finding is also the case for both experiments with $C = 10$ and $C = 20$ which indicates some degree of relation between needed communication rounds and amount of strong prosumers within the federation. This relation is further analyzed in the following part.

B. Curve Fitting

The rounds per ϕ-value to reach the target MSE from Table IV are shown in Figures 6a and 6b for $C = 10$ respective $C = 20$. In this figure, the decreasing trend with increased ϕ-value is clearly recognizable. Furthermore, it seems that the first few additional strong prosumers lead to the highest reduction in communication rounds and therefore a non-linear relation is possible. In the following, we examine two feasible functions, namely exponential and power law, defined as:



(a)

(b)

Figure 6. Curve fitting results for (a) $C = 10$ and (b) $C = 20$ with R^2-score metrics for exponential (red) and power law (blue) functions.

$$r_{\text{exp}} = a * \exp(\phi * b), \tag{9}$$

$$r_{\text{pow}} = a * \phi^{b}. \tag{10}$$

Here, the dependent variable ϕ is the strong prosumer fraction and the independent variable r is the number of communication rounds. To estimate the functions' parameters

TABLE V. Goodness of fit metrics for the curve fitted models: exponential and power law.

C	Model	$\uparrow R^2$	$\downarrow$ RMSE	$\downarrow$ AIC	$\downarrow$ BIC
10	exponential	0.86	6.38	22.53	21.75
	power law	0.99	0.65	-0.24	-1.02
20	exponential	0.64	7.99	24.79	24.01
	power law	0.91	4.09	18.12	17.33

$\Theta = [a, b]$, we employ a curve fitting based on non-linear least squares fitting approach.

$$\hat{r}_{\text{exp},10} = 63.179 \cdot \exp^{(-0.086 \cdot \phi)} \tag{11}$$

$$\hat{r}_{\text{pow},10} = 57.005 \cdot \phi^{-0.377} \tag{12}$$

$$\hat{r}_{\text{exp},20} = 55.216 \cdot \exp^{(-0.053 \cdot \phi)} \tag{13}$$

$$\hat{r}_{\text{pow},20} = 59.196 \cdot \phi^{-0.313}. \tag{14}$$

Given the observed data from Table II, the fitting process estimates a parameter vector Θ^* that maximizes the sum of squared residuals

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{N} (r_i - f(\phi, \Theta))^2. \tag{15}$$

This minimization problem is solved using the `Levenberg-Marquardt` algorithm [34], [35], which is suitable for small- to medium-sized problems with smooth, differentiable models. After solving the optimization problem, the following functions are estimated

The curve fitting is performed with the Python package `scipy.optimize`. In Figure 6, the fitted functions are shown in dashed red (power law) and blue (exponential) lines. Figure 6 suggests that a power law function describes the data points more accurate than the exponential function. But to quantify the results, *goodness of fit* metrics are applied in the following part.

C. Goodness of Fit

To evaluate the performance of the fitted models and their abilities to explain the observed data (see Table IV), we assess the goodness of fit using the *coefficient of determination*, known as R^2 metric. This metric provides a normalized measure of how much the total variance in the observed data is accounted by the respective model. Originally, the R^2 metric was developed for linear regression models, but it is widely used and applicable in non-linear context as an indicative summary statistic [36]. Formally, the R^2-score is defined as:

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(r_i - \hat{r}_i)}{\sum_{i=1}^{n}(r_i - \bar{r}_i)}, \tag{16}$$

where SSR is the residual sum of squares and SST is the total sum of squares. Thus, a R^2 score of 1 indicates a perfect fit. Conversely, an R^2 score of 0 indicates that the model performance is worse than a predicting the mean of the observed data.

Although, R^2 is useful for summarizing model fit, it should not be the only metric for model evaluation, especially in non-linear settings. Therefore, we further calculate the Root Mean Squared Error (RMSE) given as:

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{17}$$

and the Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC):

$$\text{AIC} = n \cdot \log \frac{\text{SSR}}{n} + 2k \tag{18}$$

$$\text{BIC} = n \cdot \log \frac{\text{SSR}}{n} + k \cdot \log n, \tag{19}$$

where k is the number of parameters and n the number of data points. The various metrics are listed in Table V. Based on the provided metrics, the power law models are able to describe the observed data more precise than the exponential ones. The implications are discussed in the next section.

D. Curve Fitting Implications

The observation that a power law model provides a better fit to the data than an exponential model carries important implications about the underlying system dynamics. In this case, the highest decrease in communication rounds happens within the first few additional strong prosumers. All further additions have only diminishing returns. This insight is of great interest for power grid operators and the development of distributed smart micro grids since only a few strong clients, e.g., households, are enough to accelerate training duration and therefore improve forecast accuracy for all clients within the federation.

V. Limitations

While experiment results highlight the potential of FL for the STLF problem at residential household level, several limitations need to be acknowledged. First, the provided study relied on a MLP neural network architecture with `FedAvg` as aggregation method. Although, this was a planned choice to ensure comparability with prior work and to enable implementation on distributed micro computers, it excludes more advanced model architectures, e.g., Long-Short Term Memory Neural Network (LSTM), Transformer-based models, GRU, and aggregation strategies, e.g., `FedProx`, `FedAdam`, which could yield higher forecasting accuracy. Future work should validate whether the observed power law convergence persists across those architectures. Second, the experiments were conducted with a single data set (`SmartMetersInLondon`, see Section III-A). While this data set is publicly available and also provides sufficient diversity across multiple households, it is limited to a specific geographic, temporal, and regulatory setting. Other regions may reveal different consumption patterns. Therefore, the generalization to rural grids, microgrids, or regions with higher renewable energy resources remains uncertain. Third, our proposed model restricted the input space to past consumption data without including exogenous features as weather data, calendar effects, or socio-economic indicators (see Section III-A). While the experiment design focused on unbalanced data distribution among the clients within the federation, the true forecasting potential of FL models may not be exhausted.

Despite these limitations and constraints, the findings provided by this study hold relevant implications for both research and practical application. For grid operators, the observation that only a small fraction of strong prosumers is necessary to accelerate convergence suggests that FL can be made efficient without universal data in high-resolution. This reduces infrastructure requirements and communication overhead. For prosumers, FL provides a possibility to contribute to a forecasting model without disclosing privacy-sensitive consumption data, which aligns with regulations such as the GDPR. In summary, while the presented experiments have clear methodological boundaries, they provide valuable evidence that FL can balance accuracy, efficiency, and privacy in real-world smart grid environments. The listed limitations also provide promising directions for advancing future research areas.

VI. Conclusion & Future Work

This work developed a ML-based model for the STLF problem at residential prosumer level. Given that high-resolution electricity consumption data contain behavioral information, data privacy concerns arise when transferring and processing such data. To address this, FL was incorporated as a viable approach to train ML models on distributed data without requiring direct data exchange. Three experiments were designed and conducted to evaluate the proposed FL approach. The results demonstrated that FL can achieve competitive forecasting accuracy while preserving data privacy. The trade-off between the number of learners and computational efficiency was also analyzed, along with the effects of strong and weak prosumers on training convergence and performance. Additionally, limitations of our provided work are discussed and possible solutions in future work are given.

In future work, we will focus on extending and improving the proposed FL approach. This study primarily addressed unbalanced data sets within a federation, adopting constraints such as a lightweight MLP architecture, state-of-the-art `FedAvg` weight aggregation, and the exclusion of external features. To enhance overall forecasting accuracy, these constraints should be revisited. Preliminary results indicate the utilizing more complex LSTM models and incorporating weather information can reduce forecasting errors. Additionally, this study did not explicitly implement a security layer. Future research will explore methods to ensure data privacy and prevent information leakage while integrating insights from this study. Furthermore, the potential of Transformer-based models for STLF remains an unexplored area, warranting future investigation.

Additionally, future research could explore the integration of transfer learning techniques, where forecasting knowledge gained in one region or community is transferred to another. This allows FL models trained on areas with sufficient data to support rural or emerging smart grid regions. Another promising direction is the study of incentive mechanisms for prosumers. Since FL requires active participation, especially from strong prosumers, future work should consider incentives that reward households for contributing computational resources and data.

Acknowledgment

The research in this work is supported for Alexander Wallis by the Federal Ministry of Education and Research BMBF (FKZ 03FHP212) and for Sascha Hauke by Bayern Innovativ. This is an extended version of our work presented at the *15th International Conference on Smart Grids, Green Communication and IT Energy-aware Technologies*

References

[1] A. Wallis, S. Hauke, H. Jörg, and K. Ziegler, “Federated learning for distributed load forecasting: Addressing data imbalance in smart grids,” in *The Fifteenth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, IARIA, 2025, pp. 1–2.

[2] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, et al., “Overview and importance of data quality for machine learning tasks,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3561–3562.

[3] H. Habbak, M. Mahmoud, K. Metwally, M. M. Fouda, and M. I. Ibrahem, “Load forecasting techniques and their applications in smart grids,” *Energies*, vol. 16, no. 3, p. 1480, 2023.

[4] GDPR, “General data protection regulation,” *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*, 2016.

[5] P. L. Ambassa, A. V. Kayem, S. D. Wolthusen, and C. Meinel, “Inferring private user behaviour based on information leakage,” *Smart Micro-Grid Systems Security and Privacy*, pp. 145–159, 2018.

[6] G. Wood and M. Newborough, “Dynamic energy-consumption indicators for domestic appliances: Environment, behaviour and design,” *Energy & Buildings*, vol. 35, pp. 821–841, 2003.

[7] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, “Private memoirs of a smart meter,” in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 2010, pp. 61–66.

[8] J. Chen, H. Yan, Z. Liu, M. Zhang, H. Xiong, and S. Yu, “When federated learning meets privacy-preserving computation,” *ACM Comput. Surv.*, vol. 56, no. 12, Oct. 2024, ISSN: 0360-0300. DOI: 10.1145/3679013.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54, PMLR, Apr. 2017, pp. 1273–1282.

[10] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, “A review of applications in federated learning,” *Computers & Industrial Engineering*, vol. 149, p. 106 854, 2020.

[11] P. M. Mammen, “Federated learning: Opportunities and challenges,” *arXiv preprint arXiv:2101.05428*, 2021.

[12] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, “A survey on federated learning: Challenges and applications,” *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 513–535, 2023.

[13] J. Jithish, B. Alangot, N. Mahalingam, and K. S. Yeo, “Distributed anomaly detection in smart grids: A federated learning-based approach,” *IEEE Access*, vol. 11, pp. 7157–7179, 2023.

[14] H. Liu, X. Zhang, X. Shen, and H. Sun, “A federated learning framework for smart grids: Securing power traces in collaborative learning,” *arXiv preprint arXiv:2103.11870*, 2021.

[15] X. Cheng, C. Li, and X. Liu, “A review of federated learning in energy systems,” *2022 IEEE/IAS industrial and Commercial Power System Asia (I&CPS Asia)*, pp. 2089–2095, 2022.

[16] A. GroSS, A. Lenders, F. Schwenker, D. A. Braun, and D. Fischer, “Comparison of short-term electrical load forecasting methods for different building types,” *Energy Informatics*, vol. 4, no. S3, Sep. 2021, ISSN: 2520-8942. DOI: 10.1186/s42162-021-00172-6.

[17] A. Fayyazbakhsh, T. Kienberger, and J. Vopava-Wrienz, “Comparative analysis of load profile forecasting: Lstm, svr, and ensemble approaches for singular and cumulative load categories,” *Smart Cities*, vol. 8, no. 2, p. 65, 2025.

[18] A. M. N. Ribeiro, P. R. X. do Carmo, P. T. Endo, P. Rosati, and T. Lynn, “Short-and very short-term firm-level load forecasting for warehouses: A comparison of machine learning and deep learning models,” *Energies*, vol. 15, no. 3, p. 750, 2022.

[19] K. Ullah, M. Ahsan, S. M. Hasanat, M. Haris, H. Yousaf, S. F. Raza, et al., “Short-term load forecasting: A comprehensive review and simulation study with cnn-lstm hybrids approach,” *IEEE Access*, 2024.

[20] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konený, et al., *Adaptive federated optimization*, 2021. arXiv: 2003.00295 [cs.LG].

[21] X. Wu, F. Huang, Z. Hu, and H. Huang, *Faster adaptive federated learning*, 2023. arXiv: 2212.00974 [cs.LG].

[22] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[23] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, “Scaffold: Stochastic controlled averaging for federated learning,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 5132–5143.

[24] C. Ma, J. Li, M. Ding, H. H. Yang, F. Shu, T. Q. Quek, et al., “On safeguarding privacy and security in the framework of federated learning,” *IEEE Network*, vol. 34, no. 4, pp. 242–248, 2020.

[25] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, “A survey on security and privacy of federated learning,” *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[26] M. N. Fekri, K. Grolinger, and S. Mir, “Distributed load forecasting using smart meter data: Federated learning with recurrent neural networks,” *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107 669, 2022. DOI: https://doi.org/10.1016/j.ijepes.2021.107669.

[27] A. Taïk and S. Cherkaoui, “Electrical load forecasting using edge computing and federated learning,” in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, IEEE, 2020, pp. 1–6.

[28] Y. Liu, Z. Dong, B. Liu, Y. Xu, and Z. Ding, “Fedforecast: A federated learning framework for short-term probabilistic individual load forecasting in smart grid,” *International Journal of Electrical Power & Energy Systems*, vol. 152, p. 109 172, 2023.

[29] Y. Shi and X. Xu, “Deep federated adaptation: An adaptative residential load forecasting approach with federated learning,” *Sensors*, vol. 22, no. 9, p. 3264, 2022.

[30] C. Briggs, Z. Fan, and P. Andras, “Federated learning for short-term residential load forecasting,” *IEEE Open Access Journal of Power and Energy*, vol. 9, pp. 573–583, 2022.

[31] R. Rahman, N. Kumar, and D. C. Nguyen, “Electrical load forecasting in smart grid: A personalized federated learning approach,” in *2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC)*, IEEE, 2025, pp. 1–2.

[32] *Smart meters in london*, https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london, Accessed: 2010-09-30.

[33] A. Wallis, U. Ludolfinger, S. Hauke, and M. Martens, “Towards federated short-term load forecasting,” *Internationale Energiewirtschaftstagung (IEWT 2021)*, pp. 1–9, 2021.