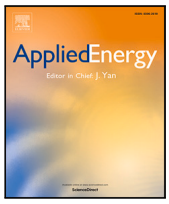
Applied Energy 384 (2025) 125419

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Personalized federated learning for household electricity load prediction with imbalanced historical data

Shibo Zhu [a,b,c], Xiaodan Shi [d,*], Huan Zhao [a,b], Yuntian Chen [c], Haoran Zhang [e], Xuan Song [f], Tianhao Wu [c], Jinyue Yan [a,b]

[a] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong
[b] International Centre of Urban Energy Nexus, The Hong Kong Polytechnic University, Hong Kong
[c] Department of Engineering, Eastern Institute for Advanced Study, China
[d] Future Energy Center, Mälardalen University, Sweden
[e] Department of Urban Planning and Design, Peking University (Shenzhen), China
[f] Department of Computer Science and Engineering, Southern University of Science and Technology, China

ARTICLE INFO

Keywords:
Load prediction
Mutual learning
Personalized federated learning
Imbalanced data

ABSTRACT

Household consumption accounts for about one-third of global electricity. Accurate results of household load prediction would help in energy management at both the building and the grid levels. Data-driven household load prediction methods have shown great advantages and potential in terms of accuracy. However, these methods still face challenges such as limited data for individual households, diversified electricity consumption behaviors, and data privacy concerns. To solve these problems, this paper proposes a personalized federated learning household load prediction framework (PF-HoLo), which allows personal models to learn collectively, leverages multisource data to capture diverse consumption behaviors, and ensures data privacy. In addition, the global encoder model and mutual learning are proposed to enhance the performance of the PF-HoLo framework considering imbalanced residential historical data. Ablation experiments results prove that the PF-HoLo framework could achieve significant improvements, with 13.41% Mean Square Error and 11.33% Mean Absolute Error, compared to traditional federated learning methods.

1. Introduction

On the consumer side of electricity, household electricity [1,2] accounts for 30%–40% of global electricity consumption [3,4], and accurate prediction results of residential electricity load brings huge benefits to household energy management [5] and grid stability [6,7]. Traditional methods use statistical and time series approaches [8–10] for the household load prediction. With the development of smart meters [11,12] in recent years, high-fidelity electricity data has become easier to achieve [13] and different deep learning (DL) models are used to predict electricity load, such as CNN [14], LSTM [15], and hybrid deep learning approach [16,17]. However, DL models require a large amount of training data and serious privacy and security issues come with it, especially at the household level. The data on electricity consumption are extremely sensitive, as they enable reconstruction of the daily behavior of users [18–20], making it difficult to share with other users to support load forecasting.

Recently, federated learning (FL) [21,22] is introduced to address data privacy problem and provide a powerful solution for electricity load prediction [19,23,24] . In a typical FL framework, the electricity consumption data of each household is saved locally, which only participates in the training of the local model, and the global model achieves better performance by integrating all local models [21].

1.1. Motivations and challenges

Although FL approaches effectively isolate user data with global models and solve privacy issues, we still face the following challenges (Fig. 1) in predicting the household electricity load.

Challenge 1: Improper User Behavior Assumption. Unlike electricity load prediction at the grid level, prediction at the household level is significantly more susceptible to individual behavior variations. In other words, the assumption for independent and identically distributed (IID) data may not hold for household load prediction. Therefore, the traditional FL methods need to be optimized for household load prediction compared to simply federating the energy usage data from different households.

* Corresponding author.
E-mail address: xiaodan.shi@mdu.se (X. Shi).

https://doi.org/10.1016/j.apenergy.2025.125419
Received 17 October 2024; Received in revised form 13 December 2024; Accepted 21 January 2025
Available online 6 February 2025

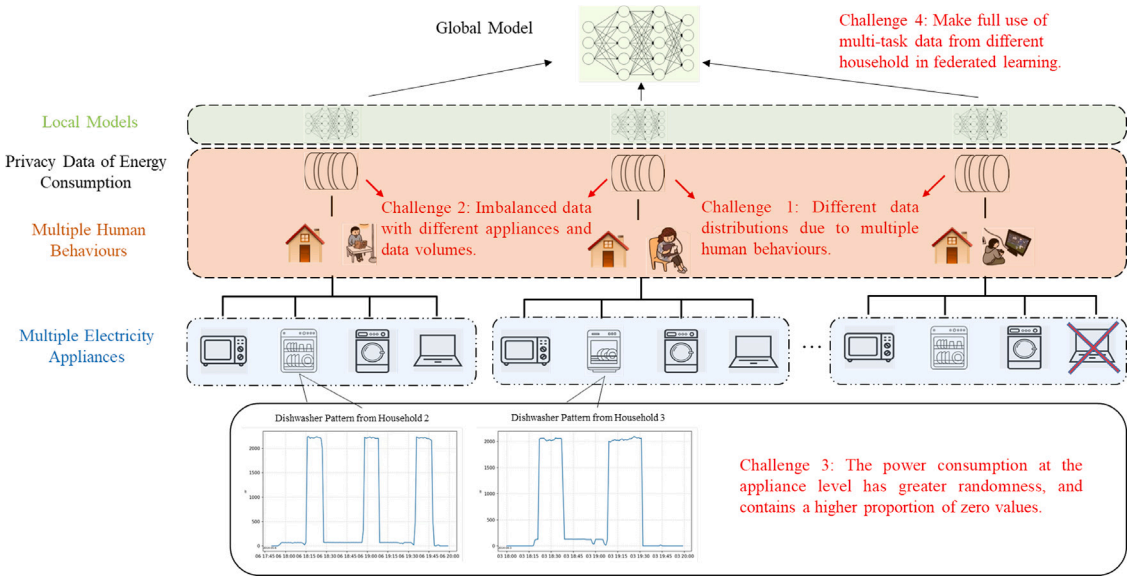


Fig. 1. Four challenges for federated learning-based household load prediction.

Challenge 2: Imbalanced Residential Data. Researchers often encounter more complex data imbalanced issues in real-world applications of federated learning for household load prediction. This is because users possess varying numbers and types of electrical appliances, even for the same kind of appliance, differences in model specifications can lead to varied electricity consumption behaviors. Furthermore, since users may start collecting data at different times, the data collected from diverse households can exhibit significant volume disparities. These resulting data imbalance phenomena place more stringent requirements on the model's recognition and generalization capabilities.

Challenge 3: Randomness and Sparseness of Appliances Data. Contrary to aggregated load data, household appliances exhibit a high degree of randomness and are typically used for short durations (Fig. 1). Consequently, the household historical load data may harder to find similar patterns and contain a higher proportion of zero values [25]. This necessitates the model to investigate deeper correlations derived from the usage patterns of various appliances, thereby imposing more stringent requirements on the model's design.

Challenge 4: Effective Local–Global Interaction Strategy. Personalized federated learning [26] presents a potential privacy-protected solution for residential load forecasting, however, it necessitates the development of an effective interaction strategy between the local household/appliance model and the overarching global model. This strategy should facilitate each household's model to leverage data from other households securely while maintaining individual household/appliance patterns, thereby enhancing its predictive accuracy.

1.2. Our contributions

To address the above-mentioned challenges, we propose a Personalized Federated Learning Household Load Prediction (PF-HoLo) framework. The framework devises FL-based strategy to empower the global model to discern and capture common patterns in electricity consumption across households, and to enable the local model to strike a balance between the idiosyncrasies of local data and the advantageous experiences gleaned from other households. The key contributions of this paper are listed as follows:

(a) The PF-HoLo framework is proposed for household load prediction with imbalanced residential data, which retains residential individualized features for each household's power consumption prediction considering the unique small data context of each household.

(b) An end-to-end federated household load prediction model is proposed with the Encoder–Decoder structure. The LSTM-based Encoder is designed to adapt to various appliance combinations and the parameters of only Encoder are shared during the learning process to maximize the shared knowledge while preserving behavior privacy.

(c) To deliver personalized load predictions for various households, the personalized loss is proposed, which utilizes the hidden state output from Encoder as the soft target for mutual learning and the model final output loss for true label learning.

(d) Experiments on appliance energy usage data from real-world households demonstrate the effectiveness of the proposed PF-HoLo framework under imbalanced data conditions. A substantial number of ablation experiments have also been conducted to validate the effectiveness of the proposed model and methods.

1.3. Organization of the paper

The remainder of the paper is organized as follows. Section 2 reviews relevant works on household electricity load prediction and federated learning. Section 3 describes our problem, framework, and methods in detail. In Section 4, we show the load prediction results and evaluate performance from quantitative and qualitative levels. Ablation experiments on crucial components are also conducted. Finally, we summarize the achievements and discuss the future work in Section 5.

2. Related work

2.1. Household electricity load prediction

The target of load prediction is to accurately forecast the electricity demand of different objects and levels, helping to continuously balance electricity supply and load demands and achieve economic, safety, and low-carbon goals [27]. According to the predicted time length, electricity load prediction can be divided into four types: very short-term load forecasting (VSTLF), short-term load forecasting (STLF), medium-term load forecasting (MTLF), and long-term load forecasting (LTLF) [28], with corresponding usage scenarios at different time granularities. In the work of this article, we mainly focus on STLF, whose prediction results can play a key guiding role in the Smart Home Energy Management System (SHMES) [27,29,30], which can be used for energy balance control and planning to improve convenience while also achieving lower electricity costs [31]. In addition, load prediction

also assists SHEMS in cost-effectively scheduling household appliances to operate on renewable energy during periods of high electricity demand [32].

While household-level consumption prediction has been extensively studied, appliance-level prediction offers more granular and actionable insights. Fine-grained data, obtained from advanced smart meters [11] or through load disaggregation techniques such as non-intrusive load monitoring (NILM) [33], can support appliance-level load forecasting and further enhance the capabilities of SHMES [27,29,30]. This approach enables precise control and optimization of individual appliances, such as scheduling dishwashers or refrigerators during off-peak hours to improve energy efficiency and maximize the use of renewable energy. Moreover, appliance-level forecasting provides valuable insights into user behavior [34], facilitating demand response programs and the development of personalized energy-saving strategies.

However, appliance-level electricity consumption prediction introduces additional challenges due to the non-stationary and stochastic nature of individual consumer behavior [25]. Compared to aggregated loads, individual household loads exhibit greater variability, making accurate prediction more complex. To address these challenges, various artificial intelligence models have been utilized, including convolutional neural networks (CNNs) [14], long short-term memory networks (LSTMs) [15], and hybrid deep learning approaches [16,17]. These methods are often enhanced through strategies such as combining models [35] or refining loss functions [25] to improve their ability to fit and predict electricity consumption data.

Despite these advancements, two critical issues remain for data-driven methods: the need for large amounts of historical data and the stringent privacy requirements of residential power consumption data. Federated learning provides a promising solution to these issues by enabling collaborative model training across distributed datasets while preserving privacy.

2.2. Federated learning

Household electricity data always faces privacy issues. Power companies generally do not disclose their customers' electricity usage information, and customers are unwilling to share their electricity usage data because it would largely reveal their behavioral information. Federated learning [21] is one way to solve this problem. In 2020, federated learning was first applied to household electricity load prediction to solve data privacy problems [36]. Fekri et al. compared two different Federated learning strategies, FedAvg and fed-sgd, and pointed out that FedAvg method has better accuracy and precision [23]. In recent years, personalized federated learning (PFL) [37] has emerged as a critical approach to addressing the challenges posed by non-IID data in FL. PFL allows for the development of models that balance shared global knowledge with local adaptability, making it particularly well-suited for household electricity consumption prediction, where data distributions vary significantly across households. Wang et al. [38] and Qu et al. [39] have proposed a personalized load prediction method for whole households by adding a local data training step after the federated learning process. However, this approach is not directly applicable to appliance-level load prediction due to the significant variability in the types and numbers of appliances across different households. Other studies have applied PFL to building-level [40] energy load forecasting through model fine-tuning [41], yet such approaches are not designed to account for the appliance-level granularity and the associated challenges of data imbalance and sparsity. Furthermore, frameworks such as Federated Mutual Learning [26], while effective in image classification tasks, have not been extensively evaluated in time-series prediction scenarios like electricity load forecasting, where temporal dependencies and irregular usage patterns present unique difficulties.

Despite these advancements, current PFL methods face limitations in handling highly imbalanced and sparse datasets, as well as in achieving a balance between global and local learning for fine-grained predictions. The PF-HoLo framework proposed in this study addresses these challenges by employing an Encoder–Decoder structure that shares only the encoder during federated learning. This design enables the global model to extract robust features from diverse households while allowing the decoder to adapt to household-specific appliance usage patterns. To further enhance personalization, a novel loss function combining soft targets (encoder-hidden states) and hard targets (model outputs) is introduced, facilitating effective mutual learning and mitigating the adverse effects of data imbalance. Unlike most existing PFL methods that focus on household-level predictions, PF-HoLo explicitly targets appliance-level load forecasting, offering fine-grained insights into energy usage and enabling more effective energy management strategies. By addressing these critical gaps, PF-HoLo provides a robust and scalable solution for personalized appliance-level load forecasting, as demonstrated in the experimental results.

3. Personalized federated learning for household load prediction

3.1. Problem definition

Different households often have different types of electrical appliances and different electricity consumption behaviors. In this problem, we will perform electricity load prediction for various electrical appliances from different households. The load prediction of household electrical appliances can be abstracted as a typical time series forecasting problem. For a specific household, the historical power consumption data of its electrical appliances are observed and used to predict the future power consumption of electrical appliances. Specifically, the historical power consumption of electrical appliances in a household can be expressed as $X = \{(p_1^t, p_2^t, \dots, p_m^t) | t \in [1, 2, \dots, T_{obs}]\}$, where p_m^t represents the power consumption of appliance m in the household at time t, T_{obs} is the length of time observed. Similarly, we define the load to be predicted as $Y = \{(p_1^t, p_2^t, \dots, p_m^t) | t \in [T_{obs} + 1, T_{obs} + 2, \dots, T_{obs} + T_{pred}]\}$, T_{obs} is the predicted time step size. The goal is to use the historical electricity load data of each household to predict their future electricity load trends.

3.2. Classical federated learning

Federated learning methods are designed to protect users' data privacy by avoiding the need to save all user data in a central database. In traditional federated learning methods, such as FedAvg, the goal is to average a global function that can achieve overall accurate performance over the distributed local data.

In the federated learning process of household load prediction, the data is distributed among various households, and each household has its own independent model with parameters ω_h to fit a local function $f_h(A_h)$, where h indicates different households. The assumption is that the electricity consumption data of each household is IID. The objective is to leverage the data from each household and obtain a model with improved performance.

To achieve this, each household in the federation agrees to update the global model using all the current models ω_h, where $h \in [1, H]$ and H indicates the number of houses in the federation. After training its own model with its own data for a specified number of epochs, the household combines its model with the models from other households to obtain a new global model G. This process allows the global model to benefit from the knowledge and insights learned from the diverse data across households, while still preserving the privacy of individual data.

$$\omega^G = \frac{1}{H} \cdot \sum_{h}^{H} \omega_h \tag{1}$$

In the described federated learning approach, the parameters ω^G of the global model G are updated by taking the average of the parameters ω_h from the sub-models in the federation. The updated global model is then distributed back to each household for the next iteration. This process is repeated for a specified number of rounds R.

During each round, the households train their local models using their own data and update their parameters based on their local optimization process. Afterward, the local models exchange their parameter updates, and the global model is updated by averaging these parameters.

By repeated iterations and parameter exchanges between the global model and the local models, the global model G gradually incorporates insights and knowledge from the diverse datasets across households. After completing all training with the specified number of rounds, the final global model G is obtained, which represents a collaborative learning result that leverages the collective intelligence of all participating households while maintaining data privacy.

3.3. PF-HoLo

3.3.1. Framework

Although the classical FL can guarantee data privacy by only sharing model parameters, it assumes that data from different families are subject to IID random variables and have a similar load pattern. However, in the real world, due to the different human behaviors and appliance types of each house, the data collected by the smart meters of each house are often not IID random variables and have different data sizes (Fig. 2). In order to solve these problems in residential load prediction, we propose a Personalized Federated Household Load Prediction Framework, **PF-HoLo**.

The overview of PF-HoLo Framework is given by (Fig. 2). Overall, it consists of two parts: **local model** and **global model**. The local model is placed in each house and consists of two models with the same network structure: the **Meme model** and the **Personal model**. Among them, the Personal model utilizes local observation data and the information transmitted by the Meme model through knowledge distillation methods [42] for training, and is responsible for outputting prediction sequences for future electricity consumption. At the same time, the Meme model conducts mutual learning based on both historical data and output from the corresponding Personal model. The difference is that the Meme model is not directly involved in forecasting future power consumption. Instead, it shares parameters with other Meme models in other families through communication to update the global model. That is to say, the Meme models in each family will have the same parameters after a certain number of training rounds due to the FedAvg operation as the training progresses. And during the training process, the parameters of the Personal models will be only updated with the mutual learning processing.

Our model draws on a work of image classification [26], which is designed to enable mutual learning between different classification tasks through the joint loss function of soft target and hard target. However, our task is a time series sequential prediction problem without corresponding probability distribution of different classes to calculate the similarity for the soft target. To address it, we further propose a simple yet useful loss function in Section 3.3.3. This improvement will help the model to better fit the distribution characteristics of local data while obtaining universal prediction features from other houses through federated learning methods, thus achieving higher prediction performance for imbalanced data conditions. We have also demonstrated this in subsequent experiments.

The overall training process of the PF-HoLo model is similar to that of the FedAvg. Specifically, during the local training process, we train both the Meme model and the Personal model mutually and enable the two models to benefit from each other's training while staying true to themselves. The specific implementation details of this part will be explained in 3.3.3. On the other hand, during the process of updating the global model, we only shared a portion of the Meme model, which will be explained in 3.3.2.

3.3.2. Meme/personal model

The Meme model and the Personal model share the same structure, which could be designed with any sequential prediction model. To enhance the model's feature extraction ability in federated learning structures with imbalanced and non-IID data conditions, the model is designed using the idea of Encoder–Decoder structure. Encoder–Decoder [43] is a type of model structure that uses a designed model to encode a sequence into a fixed-length vector representation and then uses another inverse model to decode this fixed-length vector into the target sequence. We want to integrate this technology into the PF-HoLo model to obtain a better encoder for feature extraction, while also enabling decoders in various households to better adapt to the distribution characteristics of local data.

The improved model, as shown in Fig. 3, consists of two parts: an Encoder for feature extraction and a Decoder for prediction. In the Encoder part, we choose LSTM because of its ability to capture temporal features in time series data. The Decoder consists of an LSTM model and fully connected layers.

The training process of the PF-HoLo with Encoder–Decoder structure is following. Firstly, regarding the model input, the LSTM model serving as the Encoder only processes the observed data. After the input of the observation sequence is completed, it outputs two fixed-length vectors, $h^{Enc}_{T_{obs}}$ and $c^{Enc}_{T_{obs}}$, which capture the information of T_{obs} and the preceding time series, as indicated by Eq. (2). Additionally, we consider the last observed value of each appliance in the original input data to be a crucial feature, as it provides a numerical anchor for the Decoder's predictions. We concatenate these three sets of features and use them as the initial input for the Decoder, as shown in Eq. (3).

$$h^{Enc}_{T_{obs}}, c^{Enc}_{T_{obs}} = f^{Enc}(X, h_{T_{obs}-1}, c_{T_{obs}-1}) \tag{2}$$

$$\hat{Y}, h^{Dec}_{T_{obs}}, c^{Dec}_{T_{obs}} = f^{Dec}((p_1^{T_{obs}}, p_2^{T_{obs}}, \ldots, p_m^{T_{obs}}), h^{Enc}_{T_{obs}}, c^{Enc}_{T_{obs}}) \tag{3}$$

For predicting future time steps, the Decoder takes the previous predicted value, as well as the newly generated $h^{Dec}_{T_{obs}}$ and $c^{Dec}_{T_{obs}}$ from the LSTM model in the Decoder, as inputs. This process repeats iteratively to generate the desired length of the predicted sequence $\hat{Y}$.

Besides improving the model, we have also made adjustments to the parameter-sharing strategy in federated learning for different household models. In contrast to the original model, where the entire Meme model is shared, the improved structure involves sharing only the Encoder part of the Meme model. This modification aims to enhance the encoding ability for feature extraction of federated learning, which helps to solve the imbalanced data problem by providing the feature characteristic for the few sample residences. As for the Decoder part, our goal is for it to better adapt to the power consumption characteristics of local data, so we have chosen not to allow it to directly participate in parameter sharing during federated learning. While the Encoder–Decoder structure is commonly used for time-series feature extraction, in the PF-HoLo framework, it plays a key role in addressing non-IID data. Specifically, the encoder learns shared representations across households by sharing parameters, while the decoder adapts to household-specific data distributions. This design mitigates performance degradation caused by non-IID data in federated learning.

3.3.3. Personalized loss function

To achieve personalized prediction of local electrical power consumption for each household, a configuration has been implemented wherein each household is equipped with its own Meme model and Personal model. The Meme model plays a crucial role in facilitating communication with models from other households, thereby enabling the utilization of a broader range of data. On the local front, the Personal model utilizes local data for training and achieves enhanced performance through mutual learning with the corresponding Meme models. The detailed steps involved in this process, as well as the calculation of the Loss function, are illustrated in Fig. 4.

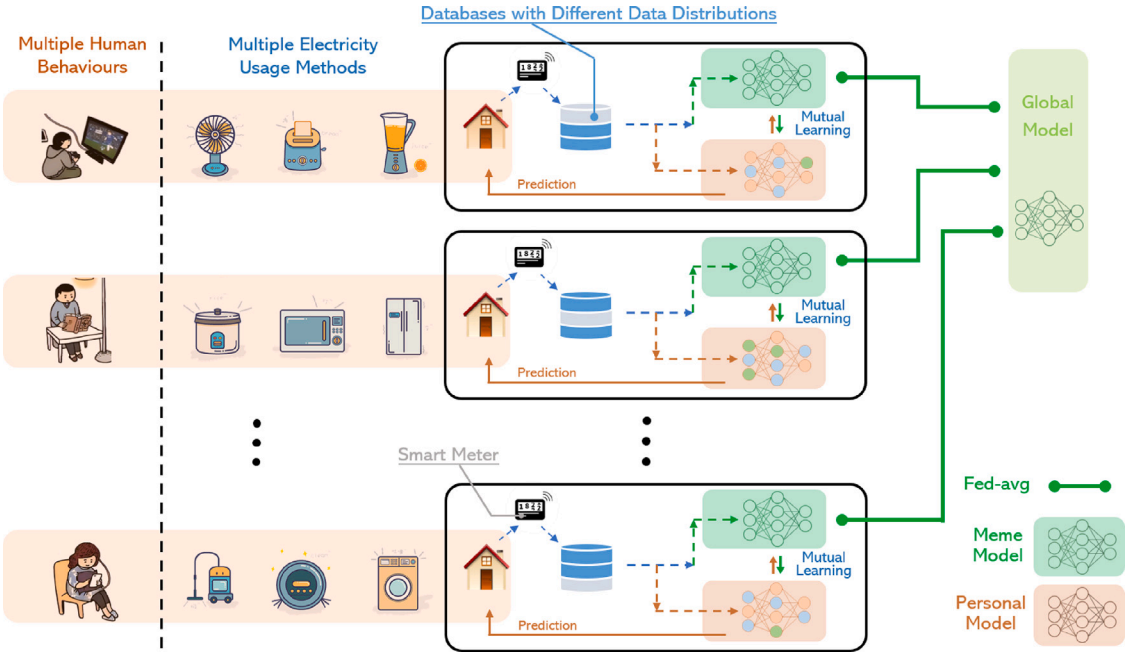


Fig. 2. The PF-HoLo framework, each household exhibits unique user behavior and a variety of electrical appliances. Each household possesses its own Meme model and Personal model, which communicate through mutual learning. The Meme model employs the FedAvg method to interact with the global model and other Meme models in households, thereby facilitating the extraction of overall electricity consumption characteristics more effectively. Conversely, the Personal model places greater emphasis on the characteristics of local data, enabling it to predict the power consumption of local appliances in a more personalized manner.

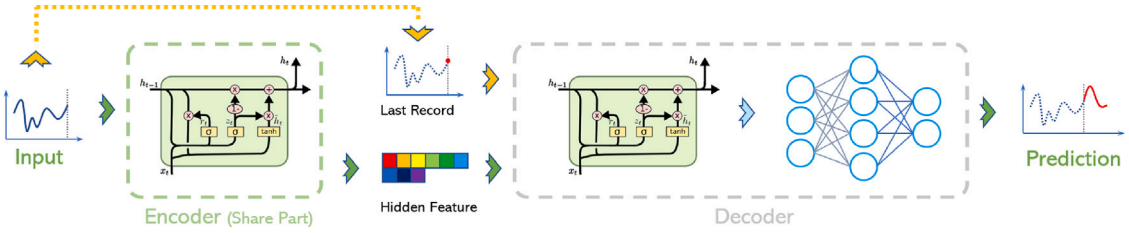


Fig. 3. Structure of Meme/Personal model.

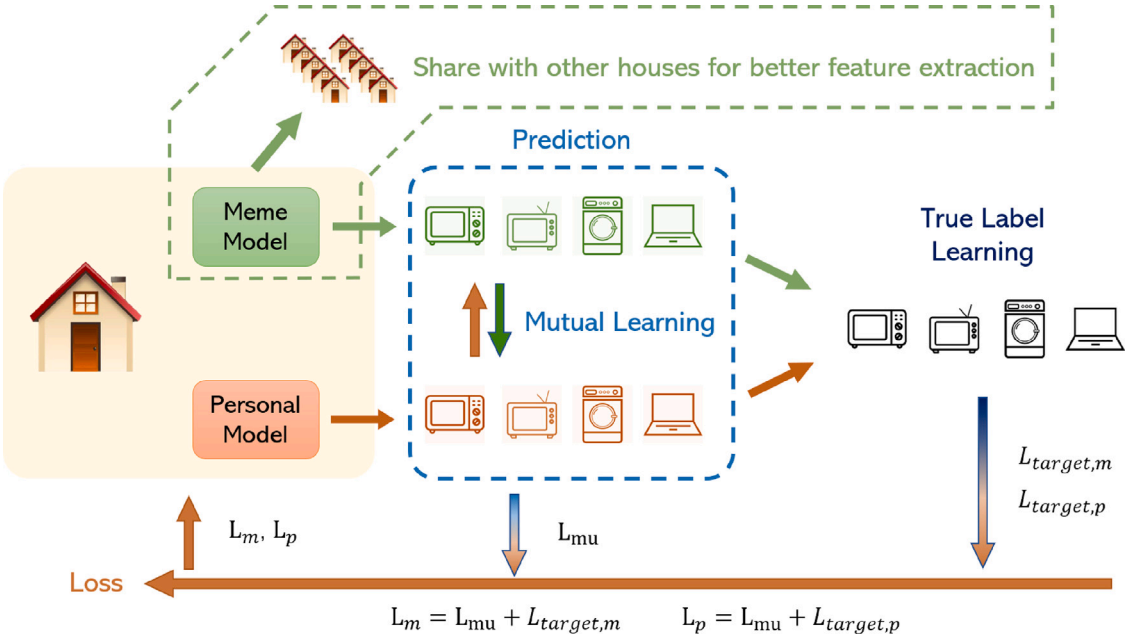


Fig. 4. Personalized loss function.

During the local training process, both the Meme model and the Personal model generate predicted electricity consumption sequences, denoted as $\hat{Y}_{Me}$ and $\hat{Y}_{Pe}$ respectively, after undergoing the process described in Eq. (3). In the process of mutual learning, the choice of knowledge distillation strategy plays a crucial role. Typically, there are two commonly used methods: hard target and soft target [44]. For the hard target, we simply need to calculate the loss using the outputs of both the Meme model and the Personal model in mutual learning, and then provide feedback to each respective model for implementation. This is relatively easy to implement and apply in temporal prediction problems. In comparison, the soft target approach in time series prediction is more challenging to implement but with better performance. For the FL image classification task [26], before the final output of the Meme and Personal models, it will generate a probability distribution through a softmax layer, which represents the probabilities of the given input belonging to different image categories. For the same input, the different probability distributions output by the previous two models can be treated as soft targets. The Kullback–Leibler (KL) divergence is then used to calculate the loss, measuring the similarity between the two models. However, in the context of time series prediction, KL divergence may not adequately capture the differences between the Meme model and the Personal model, due to the absence of probability distribution outputs in this case.

Algorithm 1: Training Process of PF-HoLo

Input : Local training round E; Max Federated training round R; Number of houses H; Number of train data N; Learning rate α; Batch size S; Observation-True Power Sequence (A_n^h, B_n^h), where $h \in [1, H], n \in [1, N]$, and X, Y with the same description in Section 3.1.

Output: Global model ω^G; Meme model ω_h^{Me} and Personal model ω_h^{Pe} for each house h;

Initialization:
$r = 1; \omega_h^{Me} = \omega^0, \omega_h^{Pe} = \omega^0, where\ (h = 1, 2, ..., H); \omega^G = \omega^0;$
for $r \le R$ **do**
  $h \leftarrow 1;$
  for $h \le H$ **do**
    Update Meme model from Global model: $\omega_h^{Me} \leftarrow \omega^G;$
    $e \leftarrow 1;$
    for $e \le E$ **do**
      Process of Mutual Train:
      $n \leftarrow 1;$
      for $n \le N$ **do**
        $\hat{Y}^{Me,h}_{n:n+S}, h^{Me,h}_{T_{obs}} \leftarrow f_{\omega_h^{Me}}(X^h_{n:n+S})$;
        $\hat{Y}^{Pe,h}_{n:n+S}, h^{Pe,h}_{T_{obs}} \leftarrow f_{\omega_h^{Pe}}(X^h_{n:n+S})$;
        $L^s_{mu} \leftarrow MSE(h^{Me,h}_{T_{obs}}, h^{Pe,h}_{T_{obs}})$;
        $L_{target,m} \leftarrow MSE(\hat{Y}^{Me,h}_{n:n+S}, Y^h_{n:n+S})$;
        $L_{target,p} \leftarrow MSE(\hat{Y}^{Pe,h}_{n:n+S}, Y^h_{n:n+S})$;
        update $\omega_h^{Me} \leftarrow \omega_h^{Me} - \alpha\nabla_{\omega_h^{Me}}(L^s_{mu} + L_{target,m})$;
        update $\omega_h^{Pe} \leftarrow \omega_h^{Pe} - \alpha\nabla_{\omega_h^{Pe}}(L^s_{mu} + L_{target,p})$;
        $n \leftarrow n + S;$
      end
      $e \leftarrow e + 1$;
    end
    $h \leftarrow h + 1$;
  end
  Communicate and update Global model:
  $\omega^G \leftarrow \frac{1}{H}\sum_{i=1}^{H} \omega_i^{Me}$;
  $r \leftarrow r + 1$;
end

To address this issue, we have proposed both hard and soft target algorithms for regression problems. Firstly, for the basic hard target approach, we can directly utilize the outputs of the Meme model $\hat{Y}_{Me}$ and Personal model $\hat{Y}_{Pe}$ to calculate $L^h_{mu} = Criterion(\hat{Y}_{Me}, \hat{Y}_{Pe})$. However, in our PF-HoLo framework, we ultimately choose to use the Soft Targets for mutual learning. In the context of electrical load prediction, this approach can provide improved predictive accuracy for the models. Specifically, with the improvement of Encoder–Decoder structure (Fig. 3), we only shared the Encoder part of the Meme model when conducting federated learning on predictive models from different households. This allows the Encoder to better leverage data from different households and obtain stronger feature extraction capabilities for the imbalanced and non-IID data conditions. To transfer this capability to the Personal model, which is trained solely on local data, we chose to use the hidden state output from the LSTM in the Encoder as the new soft target in mutual learning. We utilize this feature to compute the mutual learning loss function using Eq. (4). Two different loss calculation methods (Eqs. (5) and (6)) are used to calculate the difference between the Meme and Personal models that learn from each other from the perspective of soft targets. The final PF-HoLo model selects Mean Squared Error (MSE) as the mutual loss, and the specific comparative process is presented in the ablation experiments section.

$$L^s_{mu} = Criterion(h^M_{T_{pred}}, h^P_{T_{pred}}) \tag{4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{6}$$

Apart from the mutual learning, both Meme and Personal models also calculate the loss between their predicted values, $\hat{Y}_{Me}$ and $\hat{Y}_{Pe}$, and the actual label values. This is done as part of their respective loss functions, specified in Formulas (7) and (8). For both models, the final loss function is determined by Formulas (9) and (10), which take into account the respective true label loss and the L^s_{mu} from the mutual learning.

$$L_{target,m} = MSE(\hat{Y}_{Me}, Y) \tag{7}$$

$$L_{target,p} = MSE(\hat{Y}_{Pe}, Y) \tag{8}$$

$$L_m = L^s_{mu} + L_{target,m} \tag{9}$$

$$L_p = L^s_{mu} + L_{target,p} \tag{10}$$

Here, L_m and L_p represent the loss functions of the Meme and Personal models, respectively. $\hat{Y}_{Me}$ and $\hat{Y}_{Pe}$ denote the output sequences generated by the Meme and Personal models during the training process. Y represents the target sequences or ground truth values. Algorithm 1 provides the overall training process for the PF-HoLo framework. Firstly, we assign separate Meme models and Personal models to each household and train them only on the historical electricity load dataset of their respective households. For each household's models, a total of E*R training epochs are performed, where R represents the number of communication rounds and E represents the number of local learning epochs between two communication rounds. During the local learning phase, we calculate the individual losses and mutual loss for both the Meme model and the Personal model, and use them to update the respective models. During communication rounds, Meme models from different households share parameters with the global model, and the global model is updated. The updated global model is then transferred to each household's Meme model, replacing the previous parameters.

4. Experiments

4.1. Experimental settings

Dataset and Pre-processing Similar to some previous works [45] using federated learning, we chose the publicly available REFIT [46] dataset to test the effectiveness of our proposed model. This dataset

Table 1
Data description.

#Household	Appliances							
	Fridge	Washing machine	Dishwasher	Microwave	Kettle	Television site	Computer site	Dryer
1				x	x			
2							x	x
3							x	
4			x					x
5								
6								x
7				x			x	
8			x					
9							x	
10					x		x	x
11						x		x
12		x	x					x
13							x	x
14					x			
15				x	x			x
16			x					
17					x			
18			x				x	x
19								
20				x			x	

x represents the **absence** of records for that specific appliance in the household.

contains electrical power consumption records from 20 households in the UK over two years. Most of the data in REFIT [46] were sampled at 6–8 s intervals. In order to obtain data that can be used for model training, we first eliminated the obviously unreasonable data (the power of the distributor is greater than the total power), filled in the latest effective power with a time interval of 1 s as the minimum interval, and then resampled the data to obtain data with a time interval of 1 min. Before inputting the data into the model for training, we also performed a min–max normalization process on the data. The final predicted value of the model was inversely normalized accordingly to obtain the predicted power of the consumer in watts.

In the experiment, we selected electricity consumption data from 20 households over 13 months, from April 1, 2014, to April 30, 2015, because there were not many instances of blank data due to power outages or equipment failures during this period. In the REFIT dataset, the electricity consumption of nine types of appliances was recorded for each household, but the types of appliances recorded varied among households. We compiled and selected eight types of appliances that appeared more frequently as inputs for the model, with specific information provided in Table 1. For appliances missing in any household, we treated them as zero during model input. During the experiments, appliance-level electricity consumption predictions are obtained by modeling each appliance's power consumption as a time series. The input features include historical power consumption readings for each appliance, and the PF-HoLo framework utilizes the Encoder–Decoder structure to predict each appliance's future consumption. The prediction results for individual appliances were then aggregated to evaluate household-level performance.

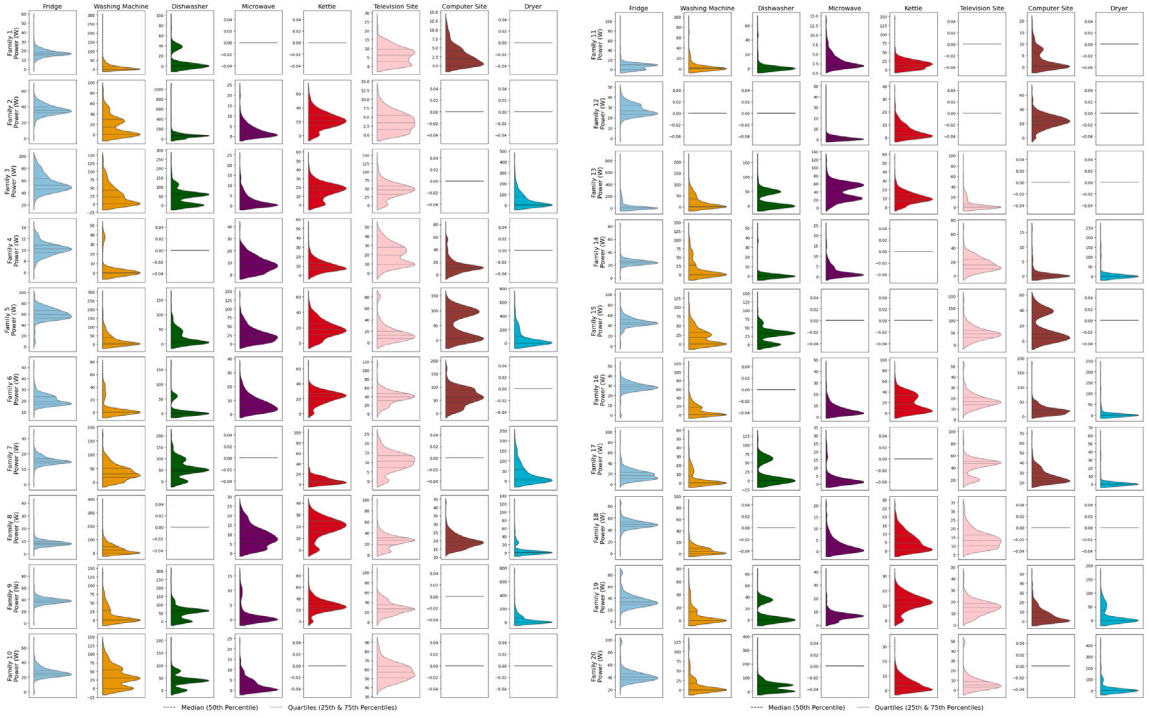
To illustrate the variability in data distribution among different households, we present single-sided violin plots of the daily average power consumption for eight types of appliances across 20 households (Fig. 5). The columns represent different appliances, while the rows show the different households. For each subgraph, the violin plot display the distribution of daily average power usage, with the median (50th percentile) indicated by a solid line, and the interquartile range (25th and 75th percentiles) visualized within the plot.

The visualizations reveal significant differences in power consumption patterns across households and appliances. For instance, the daily power consumption of refrigerators (Fridge) exhibits relatively consistent distributions across households, reflecting their continuous operation and stable consumption patterns. In contrast, appliances like dishwashers, microwaves, and kettles show highly skewed and sparse distributions, as these appliances are used intermittently and for short durations. Additionally, some households lack certain appliances entirely, as indicated by missing or flat plots.

These plots highlight the challenges posed by imbalanced and non-IID data in our study. By providing a clear visual representation of the data variability, these distributions contextualize the experimental results and demonstrate the importance of designing a federated learning framework, such as PF-HoLo, that can handle such heterogeneity effectively.

Testing strategy and Baseline In our experiments, the hardware device used is a server with Nvidia GeForce RTX 3090 GPU. All programs are written based on Python 3.8 and PyTorch 11.2.

For all experiments conducting federated learning, we set the local training round E = 5, the federated training round R = 10 (i.e., for a single model in a single family, the total number of training rounds is E * R = 50). The learning rate is 0.003. Batch size is 4096. Given that in real-world scenarios, the electricity consumption behavior and data collection duration often vary among different households. To simulate this situation, we adopted various data segmentation methods and divided these houses into three categories: Enough, Cascade, and Very Little (Table 2). ① **Enough**: Firstly, we selected six households as cases with relatively abundant data, with the proportions of data allocated to the training, validation, and test sets being 0.6, 0.2, and 0.2 of their own household data, respectively. ② **Cascade**: For another eleven households, we set the proportion of their training sets to range from 0.05 to 0.55, to mimic the diversity of real-world scenarios. ③ **Very Little**: we also explored more extreme scenarios of missing training data using the remaining three households, setting their training data proportions to 0.00625, 0.0125, and 0.025 of the total data, with the least amount of training data being approximately 2.5 days of electricity consumption records. For the latter two categories of households, we kept the proportion of validation sets at 0.2 and used all data not assigned to training or validation sets for testing to fully utilize the data. The different ratios of training to testing datasets in Table 2 were designed to simulate real-world scenarios where households have varying amounts of historical data due to differences in smart meter installation



(a) Families 1–10: Daily average power distribution of appliances

(b) Families 11–20: Daily average power distribution of appliances

Fig. 5. Daily average power distribution of household appliances across families.

Table 2
Categories and partitioning of dataset.

Categories	Cascade											Very little			Enough					
#Household	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ratio of train	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.025	0.0125	0.00625	0.6	0.6	0.6	0.6	0.6	0.6
Ratio of validation	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Ratio of test	0.75	0.7	0.65	0.6	0.55	0.5	0.45	0.4	0.35	0.3	0.25	0.775	0.7875	0.79375	0.2	0.2	0.2	0.2	0.2	0.2

periods. This setup allows us to evaluate the model's robustness under imbalanced data conditions.

In the experiment, we set the observation sequence length to 120 min and made predictions for the electrical appliance's power consumption over the next 30 min. Moreover, in subsequent experiments, we also tested its performance in predicting power consumption over 60 min, 120 min, and 180 min.

The experiments compare the performance of PF-HoLo with traditional federated learning methods such as FedAvg and FedSGD. Additionally, a centralized training model (trained without federated learning using combined data from all households) is included for comparison.

FedAvg We compared the FedAvg method to combine data from different households to test the usefulness of federated learning methods in predicting household electrical energy consumption. Specifically, in each house, there is a basic LSTM network consisting of an LSTM cell and two fully connected layers (in order to maintain consistency with the model we used in PF-HoLo). The hidden layer dimension is 128, and the dimensions of the two linear layers are (128, 32) and (32,5), respectively. The training process for FedAvg is as follows: Each family (client) is now training a certain number of rounds E on the local dataset. After the training for each family (client) is completed, the obtained model parameters will be shared with the global model, averaged, and distributed to the models in each family. This iteration involves a total of R rounds.

FedSGD FedSGD is another classical federated learning algorithm that reduces the computational overhead by transmitting gradients during training. FedSGD has an additional parameter C to control the proportion of client devices that participate in computation during each local training stage [21]. In our experiment, we set C = 1, which means that all 20 households participated in each round of federated learning. In contrast to FedAvg, the FedSGD method updates the global model with the local model gradients after each local training round, with E = 1. To ensure consistency in the number of training iterations, we increased the federated training rounds R in FedSGD to 50.

Central To evaluate the performance of centralized training, we implemented a model called Central, in which data from all households is aggregated into a single dataset for unified training. The model structure is the same as that used in FedAvg, with a basic LSTM network consisting of an LSTM cell and two fully connected layers. Unlike federated methods, Central trains directly on the combined dataset without any distributed learning process, providing a point of comparison for prediction accuracy.

4.2. Experiments result and analyze

4.2.1. Quantitative analysis

We conducted experiments using eight different electrical appliances from twenty households (Table 1). In the experiment, our model observed the electricity consumption over the past 120 min and predicted the consumption for the next 30 min. To assess the prediction accuracy, we calculated the average performance (MSE/ MAE) and used the FedAvg method as a benchmark. We then computed the

Table 3
Experimental results in MSE/MAE[a] (W^2/W) for "Cascade" data size households.

#Household		Proposed	Baseline			Ablation					
		PH-HoLo	FedSGD	FedAvg	Central	EnDe FedMu (Soft MAE)	EnDe FedMu (Hard)	EnDe FedAvg	FedMu (Soft MSE)	FedMu (Soft MAE)	FedMu (Hard)
Cascade	1	4049.20/8.71	4620.46/8.70	4656.41/8.97	4915.35/10.17	4330.59/8.62	4352.85/8.04	4377.39/8.05	4562.65/8.68	4571.02/10.18	4678.14/7.94
	2	14 498.87/21.64	15 364.92/22.32	14 962.57/22.22	15 209.13/20.54	14 865.89/21.75	15 068.94/21.92	14 408.22/20.23	15 580.24/23.50	15 043.50/22.43	16 298.46/23.73
	3	14 172.31/24.68	20 168.45/30.75	19 640.60/31.95	20 188.70/28.85	15 628.21/24.79	15 014.12/25.55	17 665.36/28.30	17 635.45/27.49	17 767.30/28.49	19 922.85/31.90
	4	4811.38/9.86	4984.01/11.15	5018.88/11.35	5068.98/12.05	4798.60/9.99	4773.39/9.51	4867.52/9.95	4754.62/10.32	4919.73/10.79	4947.56/10.50
	5	15 510.91/25.88	21 831.14/32.86	21 086.58/31.19	21 065.55/32.29	16 145.33/26.67	15 805.56/25.49	18 240.77/28.38	16 700.64/28.13	17 435.70/30.33	17 864.79/28.10
	6	8899.69/16.03	9029.00/17.43	9123.20/17.11	9198.61/16.91	8867.23/15.89	8811.27/15.95	8794.62/15.53	8973.52/17.15	9182.13/16.88	9296.20/18.45
	7	21 602.01/26.42	26 367.54/29.19	25 702.21/30.28	25 647.79/27.07	21 440.61/25.39	21 617.84/26.96	23 241.11/28.05	22 049.12/29.52	22 162.27/26.97	23 985.81/27.35
	8	12 282.34/16.02	13 418.69/17.72	13 255.46/18.19	13 404.67/16.01	12 267.04/15.65	12 417.92/16.95	12 809.78/16.47	12 587.16/17.03	12 803.31/17.33	13 606.73/18.88
	9	16 906.18/23.09	18 307.03/25.49	18 258.12/25.05	18 957.60/24.28	16 782.69/23.27	17 271.66/23.15	17 602.55/23.75	17 487.32/25.06	17 469.25/24.87	18 464.12/26.54
	10	6223.97/17.80	7916.32/19.40	7629.49/18.91	7306.25/17.22	5901.80/16.65	6297.94/18.87	6141.25/16.93	7170.76/20.61	5971.50/18.84	8430.12/20.68
	11	5328.42/7.84	5885.91/9.65	5854.62/8.95	6137.99/10.86	5465.44/8.39	5396.85/7.95	5610.83/7.94	5645.80/8.96	5553.44/9.12	5678.83/8.27
AVG		11 298.66/18.00	13 444.86/20.42	13 198.92/20.38	13 372.78/19.66	11 499.40/17.92	11 529.85/18.21	12 159.95/18.51	12 104.30/19.68	12 079.92/19.66	13 015.78/20.21
Compare to FedAvg		**14.40%**/11.69%	−1.86%/−0.22%	///	−1.32%/3.54%	12.88%/**12.09%**	12.65%/10.64%	7.87%/9.20%	8.29%/3.45%	8.48%/3.55%	1.39%/0.82%

[a] A lower MSE and MAE indicate superior model performance.
[b] The term **bold** denotes the superior performance among comparable entities.

Table 4
Experimental results in MSE/MAE[a] (W^2/W) for "Very Little" data size households.

#Household		Proposed	Baseline			Ablation					
		PH-HoLo	FedSGD	FedAvg	Central	EnDe FedMu (Soft MAE)	EnDe FedMu (Hard)	En-De FedAvg	FedMu (Soft MSE)	FedMu (Soft MAE)	FedMu (Hard)
Very little	12	11 020.23/14.58	11 231.10/14.48	11 212.31/15.22	11 469.42/14.19	11 121.89/14.52	10 758.43/14.29	11 152.22/14.50	11 145.25/15.27	11 136.31/14.44	11 191.46/14.46
	13	10 438.21/15.01	13 417.53/19.39	13 083.98/17.54	14 005.00/18.57	11 602.83/16.63	10 940.33/14.87	12 376.13/15.85	12 824.25/16.19	12 263.11/16.60	12 418.08/16.89
	14	2628.17/7.24	3612.97/9.09	3505.80/9.82	3653.90/9.45	2700.57/7.89	2720.28/7.51	3081.63/8.00	3330.23/8.79	3146.48/8.49	3141.37/8.78
AVG		8028.87/12.28	9420.53/14.32	9267.36/14.19	9709.44/14.07	8475.10/13.01	8139.68/12.22	8869.99/12.79	9099.91/13.42	8848.63/13.18	8916.97/13.38
Compare to FedAvg		**13.36%**/13.50%	−1.65%/−0.89%	///	−4.77%/0.90%	8.55%/8.30%	12.17%/**13.89%**	4.29%/9.91%	1.81%/5.47%	4.52%/7.15%	3.78%/5.74%

[a] A lower MSE and MAE indicate superior model performance.
[b] The term **bold** denotes the superior performance among comparable entities.

percentage difference in performance between the models. Table 3, 4 and 5 respectively present the results of predicted electricity consumption for three types of households: Enough, Cascade, and Very Little. These tables also show the results of our ablation experiments, which will be discussed in detail in Section 4.3. From these tables, it is evident that our proposed method consistently achieved better prediction performance across most cases. In the comparison of baselines, we observed that the overall prediction performance of FedAvg, FedSGD, and the Central model does not differ significantly, indicating that federated learning can serve as a viable alternative to centralized learning in this scenario. However, the centralized model generally achieves better performance in terms of MAE for most households, while federated learning models tend to perform better on the MSE metric. This discrepancy may be attributed to the centralized model's reliance on global data distributions, which can result in the neglect of important local information specific to individual households. In contrast, our proposed model effectively balances global information and local features, achieving significant improvements on both metrics.

Compared to centralized model and traditional federated learning approaches, our proposed PF-HoLo model can more effectively utilize data from different households, which have varying electrical appliances. This allows the model to make personalized predictions about local electricity consumption and achieve better prediction accuracy when faced with households having different electrical devices and data sizes. On average, our model improved prediction accuracy by approximately 13.41% and 11.33% in MSE and MAE, respectively, compared to the FedAvg method (Fig. 10).

In addition, through cross-comparison of the model performance in three different categories of households with varying training data sizes, we observed that the proposed PF-HoLo model exhibited a greater relative improvement in prediction accuracy for the Cascade and Very Little categories. This suggests that our model provides a better enhancement in user experience for federated learning scenarios where households have imbalanced historical data. The personalized federated mutual learning approach can leverage the characteristics of local data to achieve better prediction results even with limited user data. We also presented the performance of other model combinations in household load forecasting, which will be discussed in detail in the ablation experiments section.

To further evaluate the performance of our proposed PF-HoLo model, we analyzed the R^2 scores, which measure the proportion of variance in the target variable that is explained by the model. Fig. 6(a) presents the R^2 scores achieved by PF-HoLo across different appliances and households, while Fig. 6(b) illustrates the improvement in R^2 scores compared to the FedAvg method.

As shown in Fig. 6(a), the PF-HoLo model achieves consistently high R^2 scores for certain appliances, such as dishwashers and televisions. Dishwashers typically exhibit regular and periodic usage patterns, making their consumption easier to predict. Similarly, televisions often follow predictable patterns based on user habits, contributing to higher R^2 scores. In contrast, for appliances with more irregular usage, such as microwaves and kettles, the R^2 scores are relatively lower. This reflects the inherent challenge of predicting the consumption of appliances with highly stochastic and short-duration usage. Additionally, refrigerators and washing machines also show relatively low R^2 scores despite their regular usage patterns. The low R^2 scores for refrigerators can be attributed to their stable and nearly constant power consumption with minimal variation, which inherently reduces the total variance in the data, thereby limiting the achievable R^2 score. For washing machines, the intermittency and multiple operational phases (e.g., washing, rinsing, spinning) introduce complexity in power consumption patterns, making accurate predictions more difficult.

Fig. 6(b) highlights the improvement of PF-HoLo over FedAvg in terms of R^2 scores. On average, PF-HoLo demonstrates noticeable improvements across most households and appliances, with particularly significant gains for appliances like computer sites. These improvements underscore the ability of PF-HoLo to better capture localized and appliance-specific consumption patterns, benefiting from its personalized federated learning framework.

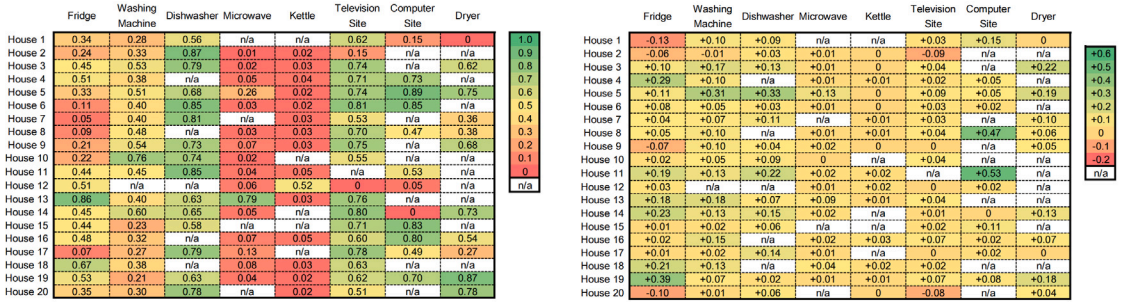
Overall, the R^2 analysis further validates the effectiveness of PF-HoLo in balancing global and local information, enabling it to outperform traditional federated learning methods in capturing the variability of household electricity consumption.

To more vividly illustrate the superiority of our proposed method relative to the baseline model, we graphically represent our proposed model to compare its relative accuracy with the FedAvg approach in predicting performance in different households and appliances, as depicted in Fig. 7. Here, due to the variations in appliance ownership

Table 5
Experimental results in MSE/MAE[a] (W^2/W) for "Enough" data size households.

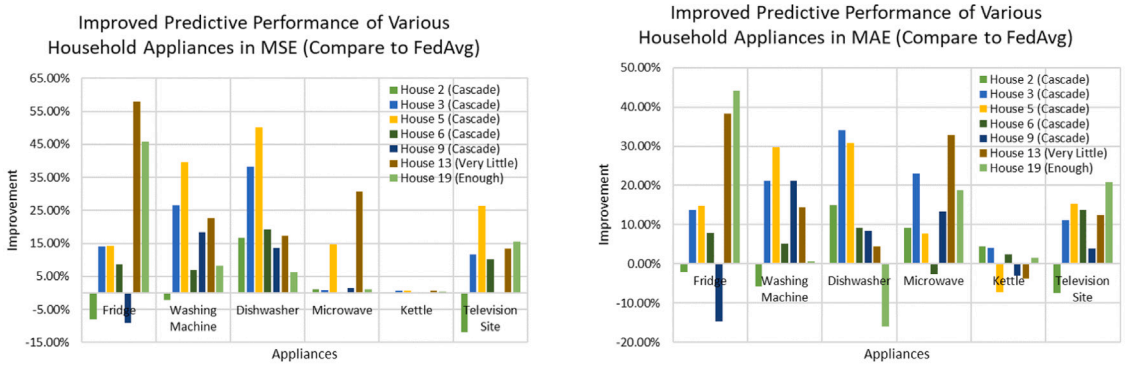
#Household		Proposed	Baseline			Ablation					
		PH-HoLo	FedSGD	FedAvg	Central	EnDe FedMu (Soft MAE)	EnDe FedMu (Hard)	En-De FedAvg	FedMu (Soft MSE)	FedMu (Soft MAE)	FedMu (Hard)
Enough	15	10 703.02/21.34	11 600.06/23.57	11 559.34/22.32	11 618.38/20.07	10 671.77/21.33	10 807.00/21.33	10 958.53/20.88	11 185.68/21.81	11 345.56/22.67	11 540.90/22.41
	16	8572.08/13.53	9341.88/15.41	9193.96/14.13	9513.79/15.18	8669.10/13.23	8671.50/12.82	8956.89/13.64	8780.86/13.71	8804.56/13.06	9027.04/14.52
	17	3334.68/9.72	4340.34/11.35	4445.19/10.84	4187.14/11.01	3421.47/9.99	3527.26/9.53	3639.50/9.57	4343.44/11.02	4111.14/11.23	5624.03/13.77
	18	3548.36/9.22	3922.77/12.25	3838.20/11.03	4108.40/15.12	3675.16/10.37	3470.43/9.09	3734.17/10.05	3709.19/10.39	3693.64/10.29	3859.55/11.26
	19	6881.53/12.07	7816.55/15.27	7749.53/15.24	8391.62/14.95	7100.80/12.96	6789.58/11.91	7208.26/12.10	7258.31/13.01	7297.48/15.27	7135.61/12.15
	20	6569.91/14.75	7797.58/16.58	7311.28/15.41	8035.43/17.62	6345.65/15.11	6370.87/13.42	6822.33/14.95	6511.84/15.31	6857.78/16.76	7267.07/15.84
AVG		6601.60/13.44	7469.86/15.74	7349.58/14.83	7642.46/15.66	6647.32/13.83	6606.11/13.02	6886.61/13.53	6964.89/14.21	7018.36/14.88	7409.03/14.99
Compare to FedAvg		**10.18%**/9.37%	−1.64%/−6.13%	///	−3.98%/−5.60%	9.56%/6.72%	10.12%/**12.22%**	6.30%/8.74%	5.23%/4.17%	4.51%/−0.34%	−0.81%/−1.11%

[a] A lower MSE and MAE indicate superior model performance.
[b] The term **bold** denotes the superior performance among comparable entities.



(a) R^2 scores for PF-HoLo across households (b) R^2 score improvements of PF-HoLo over FedAvg

Fig. 6. R^2 Scores and improvements for appliance-level prediction.



(a) MSE (b) MAE

Fig. 7. Performance of PF-HoLo in appliance level (Compare to FedAvg).

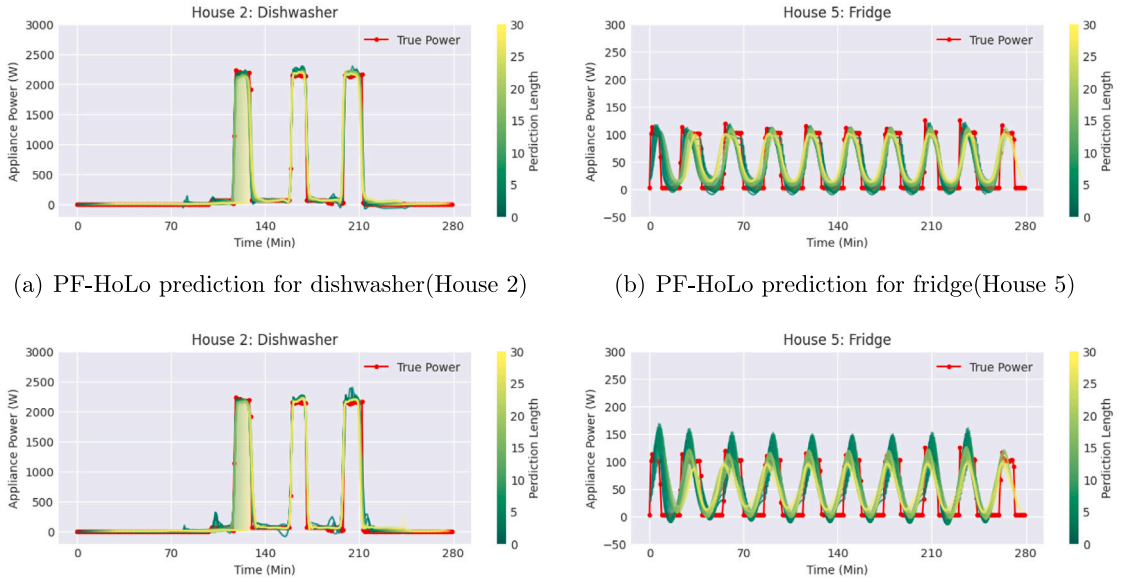
among different households (Table 1), we selected six households with five identical types of appliances. Among these households, four are from the Cascade category, while the remaining two are from the Very Little and Enough categories, respectively.

From Fig. 7(a), when examining the consumer electricity consumption prediction across different households, it can be observed that the improvements of the proposed model mainly focus on these four types of appliances: fridge, washing machine, dishwasher, and television site. The accuracy improvement for these appliances is significant compared to the FedAvg method. However, for the other two types of appliances, microwave ovens and hot water kettles, the accuracy improvement is relatively limited.

One possible reason for this discrepancy is the sampling frequency used in the experiment. The data were resampled with a time interval of 1 min. However, microwave ovens and hot water kettles often have short durations of use, making it difficult for the model to capture their electrical characteristics accurately within such a coarse-grained time interval.

In contrast, the first four types of appliances typically have longer operating periods. For example, a fridge may have a clear cycle of cooling, reaching the target temperature, going into standby mode, and then operating again to cool. Similarly, washing machines, dishwashers, and televisions typically operate for more than 30 min, providing the model with more opportunities to capture their electricity consumption characteristics for better prediction. A similar trend can be observed when analyzing the effectiveness based on the MAE evaluation metric (Fig. 7(b)). However, the MAE metric specifically shows better performance for the microwave oven in our method.

Overall, Fig. 7 highlights the effectiveness of the proposed model in improving the prediction accuracy of consumer power consumption, particularly for certain types of appliances. The results indicate that the model performs better for appliances with longer operating durations, while the accuracy improvement for appliances with shorter durations is relatively smaller. Furthermore, for households with different electricity usage habits, the prediction accuracy of our proposed method is overall superior to the FedAvg method in terms of electricity consumption forecasting.



(a) PF-HoLo prediction for dishwasher(House 2)

(b) PF-HoLo prediction for fridge(House 5)

(c) FedAvg prediction for dishwasher(House 2)

(d) FedAvg prediction for fridge(House 5)

Fig. 8. Visualization of electrical load prediction.

4.2.2. Qualitative analysis

The partial results of our experiment are visualized in Fig. 8. Within Fig. 8, we present the performance of the proposed model on two distinct electrical appliances across two households, as depicted in Figs. 8(a) and 8(b). These results are juxtaposed with the predictions from the FedAvg method in Figs. 8(c) and 8(d). In each subplot, the x-axis signifies relative time and the y-axis denotes the power consumption of the electrical appliances.

In Fig. 8(a), a red dotted line represents the actual power utilization of the dishwasher in Household 2. The 30 gradient curves, transitioning from green to yellow, illustrate the model's predictions for power consumption for the upcoming t-th minute, where $t \in [1, 30]$. For instance, a yellow point on the x-axis marked at 140th minute corresponds to the prediction for power consumption 30th minute into the future, which is based on the genuine power consumption represented by the red dotted line from −10 min to 110 min (the curve before the 0-min mark is not shown in the figure). In this context, our observation period has been set to 120 min.

Fig. 8(a) depicts a typical working scenario of a dishwasher in Household 2. By observing the real load curve, we can see that the dishwasher starts operating after 70-min mark, goes through a preheating phase, experiences three power peaks with a maximum power of around 2200 W, and then the power consumption drops to 0 as the operation concludes. Comparing the prediction results of the proposed PF-HoLo (Fig. 8(a)) and FedAvg (Fig. 8(c)) models, we can observe significant improvements in the prediction of the startup phase and power peaks in this example. During the startup phase, FedAvg incorrectly predicts a power consumption of approximately 400 W in the short-term forecast, while PF-HoLo accurately predicts the power consumption at the startup of the dishwasher. Furthermore, during the third power peak, the power consumption prediction from FedAvg is notably more unstable, forecasting a short-term power consumption of around 2400 W. In contrast, PF-HoLo does not exhibit this issue. In terms of long-term forecasting, we notice that the FedAvg method incorrectly predicts a small peak of approximately 300 W around 240-min mark, whereas PF-HoLo does not exhibit this problem.

Additionally, we visualized the predictive performance of our proposed model in comparison with the baseline for a refrigerator in Household 5 (Figs. 8(b) and 8(d)). Compared to a dishwasher, the power consumption of a fridge exhibits stronger periodicity and has a narrower power range. Our model shows significant improvement in short-term power consumption prediction (represented in green), accurately describing the power variations without erroneously predicting the power at the peak of 150 W, as was the case with FedAvg. Furthermore, for longer-range predictions (indicated in yellow), the proposed model demonstrates better accuracy and consistency. We observe that the yellow curve in Fig. 5(b) consistently depicts the periodic trend of refrigerator power consumption, whereas, under the FedAvg method, the yellow curve representing longer-range prediction exhibits noticeable stratification and a less accurate depiction of the periodic trend compared to PF-HoLo.

4.2.3. Prediction of different length

Fig. 9 presents a performance comparison between the proposed model and the baseline model, considering an observation step of 120 and prediction steps of 30, 60, 120, and 180 minutes, respectively, when the data is sampled at a 1-min interval. The evaluation is based on two metrics: average squared error (Fig. 9(a)) and average absolute error (Fig. 9(b)). Smaller values for both metrics indicate better performance.

The results clearly demonstrate that the proposed method, specifically the PF-HoLo model and its fine-tuning version, outperforms the baseline method in terms of prediction accuracy under different prediction step sizes. The improvement in accuracy is significant.

Moreover, it is observed that as the prediction step size increases, the error between the predicted results of all methods (including the proposed and baseline methods) and the true power gradually increases. This is expected because with larger prediction steps, the uncertainty related to the future condition of electrical appliances increases, making accurate predictions more challenging.

However, even with the increasing difficulty in predicting larger steps, the proposed method still demonstrates superior performance compared to the baseline method. This indicates that the proposed method maintains better prediction performance even within a 180-min prediction window.

4.3. Ablation experiment

4.3.1. Ablation experiment setting

As stated in Section 3.3.3, in mutual learning, the choice of knowledge distillation strategy is important to the model's performance.

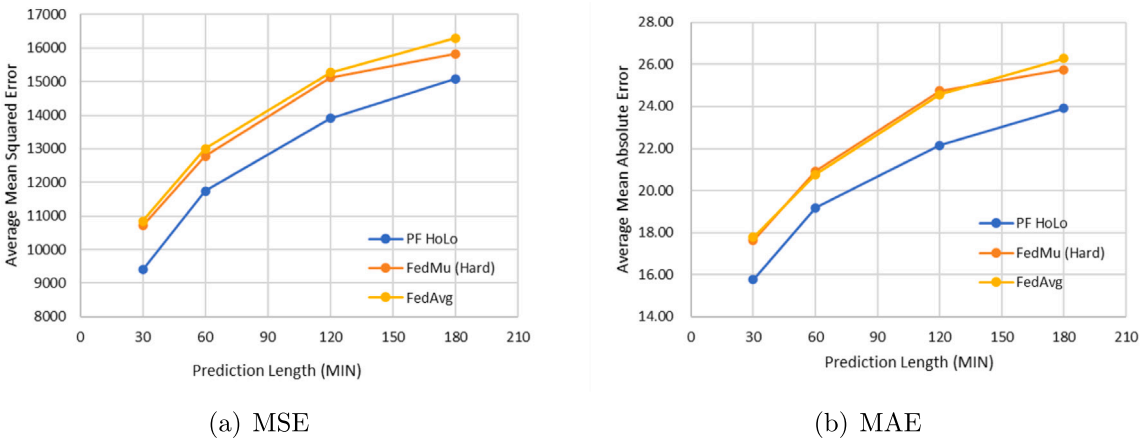


Fig. 9. Performance of models with different predict time length.

Hence, we experimented with different combinations of knowledge distillation targets and loss functions by ablation settings: ❶**FedMu (Soft MAE)**, it shares the same model architecture with PF-HoLo and uses the soft target for mutual learning. And we change the criterion in Eq. (4) from MSE to MAE, to verify the effect of different criteria of soft targets; ❷**EnDe FedMu (Hard)**, it also shares the same model architecture with PF-HoLo, but only uses the hard target as mutual loss. We directly use the outputs of its own Meme model $\hat{Y}_{Me}$ and Personal model $\hat{Y}_{Pe}$ as the hard target to calculate mutual loss $L^{h}_{mu} = MSE(\hat{Y}_{Me}, \hat{Y}_{Pe})$; ❸**EnDe FedAvg**. To further verify the effectiveness of mutual learning, we remove the mutual learning component from PF-HoLo and use the Fed Avg method as a replacement.

We further conducted similar ablation experiments, building on the removal of the Encoder–Decoder structure, to verify the effectiveness of this structure and the performance of different target selections in this case: ❹**FedMu (Soft MSE)**, it is similar to the PF-HoLo, but we removed the Encoder–Decoder structure to verify its effectiveness. Due to the absence of this structure, we made a slight adjustment to the choice of the soft target. We selected the hidden state of the LSTM at the last time step during prediction as the soft target, denoted as $h^{M}_{T_{obs}+T_{pred}}$ and $h^{P}_{T_{obs}+T_{pred}}$, respectively, since they capture all the information up to and including that time point. Then, the mutual loss L^{s2}_{mu} was computed using Eq. (11) with MSE; ❺**FedMu (Soft MAE)**, shares the same architecture with the setting ❹, but the criterion of soft target was changed to MAE; ❻**FedMu (Hard)**, it only uses hard target as the mutual loss, but compared to ablation setting ❷, we further removed the Encoder–Decoder structure to verify the role of this structure.

$$L^{s2}_{mu} = Criterion(h^{M}_{T_{obs}+T_{pred}}, h^{P}_{T_{obs}+T_{pred}}) \tag{11}$$

Overall, by comparing the ablation setting group ❶, ❷, ❸ and ❹, ❺, ❻, we can verify the effectiveness of the Encoder–Decoder structure in our proposed model. At the same time, by combining these ablation experiments and the baseline experiments, we can also comparatively verify the advantages brought by federated mutual learning and compare the impact of different target selections in mutual learning on model accuracy.

4.3.2. Ablation experiment result

Fig. 10 presents the performance results of models employing previously mentioned soft targets, hard targets, and various loss calculation methods in power consumption prediction relative to the FedAvg method. First, by comparing the federated mutual learning model with the original model, we observed that different mutual learning targets led to improvements in prediction performance compared to the FedAvg method. However, the improvements were more pronounced when soft targets were used. This could be attributed to the fact that using hidden states as soft targets, instead of directly using the final prediction output of the model as hard targets, introduced more information and aided the Personal models in achieving better prediction results. Second, we found that the introduction of the Encoder–Decoder structure greatly enhanced prediction accuracy. Even without mutual learning, simply adapting the FedAvg method to the Encoder–Decoder structure achieved similar improvements as using soft targets in terms of MSE metric, and even more significant improvements in terms of MAE metric. Nevertheless, the additional benefits brought by mutual learning were still evident. Under the Encoder–Decoder structure, all federated learning methods achieved significant improvements in prediction accuracy. Among them, our proposed PF-HoLo method performed better in terms of the composite metrics of MSE and MAE, exhibiting respective improvements of 13.41% and 11.33% compared to the FedAvg method.

5. Conclusion

This paper studies the problem of personalized household load prediction, revealing a series of challenges and providing a new solution.

First, the federated learning household load prediction framework, PF-HoLo, is proposed to solve the short-term load forecasting problem, which to a large extent ensures the privacy of users and increases the forecasting accuracy. In the context of the popularity of smart meters, it is possible to promote electricity load prediction technology on a large scale.

Secondly, in response to the common problem of non-IID data in federated learning, we introduce the Encoder–Decoder structure and personalized loss function. These adaptations promote the PF-HoLo framework suitable for predicting the power consumption of multiple households and appliances with imbalanced data conditions.

Finally, we conducted extensive ablation experiments on the PF-HoLo, demonstrating the effectiveness of the Encoder–Decoder structure and the soft target knowledge distillation method in this framework. Our proposed PF-HoLo framework demonstrated an overall performance improvement compared to the traditional FedAvg method with various electrical appliances and household data size conditions. Specifically, it achieved a 13.41% improvement in terms of the MSE metric and an 11.33% improvement in terms of the MAE metric.

In future work, we plan to address the limitation posed by the absence of appliance-specific sensors in many households by integrating non-intrusive load monitoring (NILM) techniques. These techniques can disaggregate total household electricity consumption into appliance-level usage, which can then serve as input to our model, enabling the deployment of the framework in sensor-free environments.

Furthermore, appliance-level predictions offer notable advantages, such as enabling integration with demand response programs and

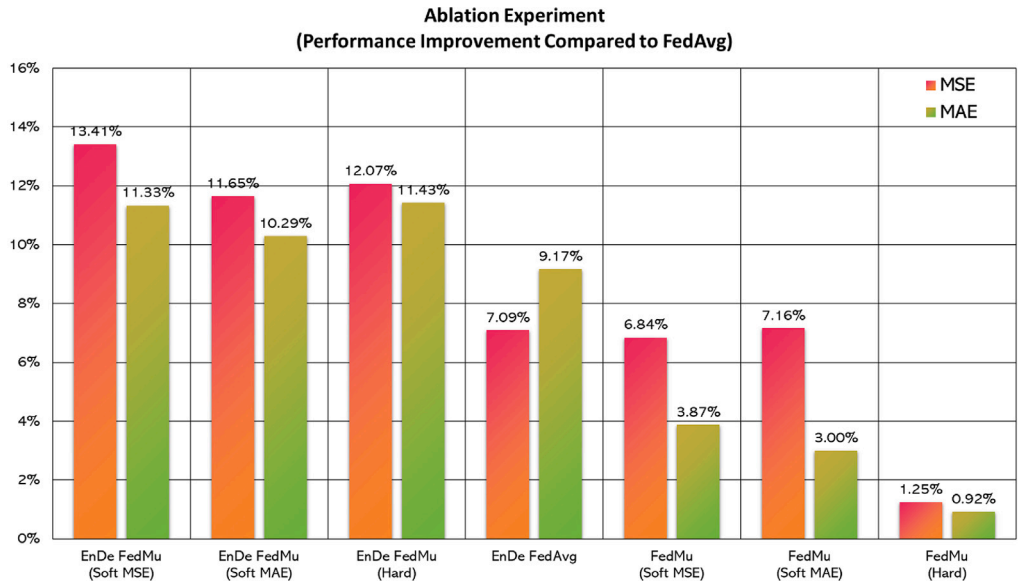


Fig. 10. Ablation experiment results (Compare to FedAvg).

providing more granular insights into user behavior, which are particularly valuable for utilities and policymakers. However, household-level predictions, while reducing computational costs, lack the adaptability and personalization benefits of appliance-level approaches. To better understand this trade-off, we plan to include a systematic comparison of household-level versus appliance-level predictions as a key focus of our future work.

CRediT authorship contribution statement

Shibo Zhu: Writing – original draft, Visualization, Software, Methodology, Data curation. **Xiaodan Shi:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Huan Zhao:** Writing – review & editing. **Yuntian Chen:** Resources. **Haoran Zhang:** Resources. **Xuan Song:** Resources. **Tianhao Wu:** Resources. **Jinyue Yan:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the RISUD Interdisciplinary Research Scheme (1-BBWW), International Research Centre of Urban Energy Nexus (P0047700). The authors also thank the EIT High Performance Computing Platform for providing computational resources for this project.

Data availability

The authors do not have permission to share data.

References

[1] Stute Judith, Pelka Sabine, Kühnbach Matthias, Klobasa Marian. Assessing the conditions for economic viability of dynamic electricity retail tariffs for households. Adv Appl Energy 2024;14:100174.
[2] Xiang Xiwang, Zhou Nan, Ma Minda, Feng Wei, Yan Ran. Global transition of operational carbon in residential buildings since the millennium. Adv Appl Energy 2023;11:100145.
[3] Haider Haider Tarish, See Ong Hang, Elmenreich Wilfried. A review of residential demand response of smart grid. Renew Sustain Energy Rev 2016;59:166–78.
[4] Haq Ejaz Ul, Lyu Xue, Jia Youwei, Hua Mengyuan, Ahmad Fiaz. Forecasting household electric appliances consumption and peak demand based on hybrid machine learning approach. Energy Rep 2020;6:1099–105.
[5] Khan Zulfiqar Ahmad, Ullah Amin, Haq Ijaz Ul, Hamdy Mohamed, Mauro Gerardo Maria, Muhammad Khan, Hijji Mohammad, Baik Sung Wook. Efficient short-term electricity load forecasting for effective energy management. Sustain Energy Technol Assess 2022;53:102337.
[6] Zhu Jizhong, Dong Hanjiang, Zheng Weiye, Li Shenglin, Huang Yanting, Xi Lei. Review and prospect of data-driven techniques for load forecasting in integrated energy systems. Appl Energy 2022;321:119269.
[7] Du Ying, Liu Yadong, Yan Yingjie, Fang Jian, Jiang Xiuchen. Risk management of weather-related failures in distribution systems based on interpretable extra-trees. J Mod Power Syst Clean Energy 2023;11(6):1868–77.
[8] Wang Yi, Gan Dahua, Zhang Ning, Xie Le, Kang Chongqing. Feature selection for probabilistic load forecasting via sparse penalized quantile regression. J Mod Power Syst Clean Energy 2019;7(5):1200–9.
[9] Kipping A, Trømborg E. Modeling and disaggregating hourly electricity consumption in norwegian dwellings based on smart meter data. Energy Build 2016;118:350–69.
[10] Laouafi Abderrezak, Mordjaoui Mourad, Laouafi Farida, Boukelia Taqiy Eddine. Daily peak electricity demand forecasting based on an adaptive hybrid two-stage methodology. Int J Electr Power Energy Syst 2016;77:136–44.
[11] Wang Fei, Lu Xiaoxing, Chang Xiqiang, Cao Xin, Yan Siqing, Li Kangping, Duić Neven, Shafie-Khah Miadreza, Catalão João PS. Household profile identification for behavioral demand response: A semi-supervised learning approach using smart meter data. Energy 2022;238:121728.
[12] Tang Wenjun, Wang Hao, Lee Xian-Long, Yang Hong-Tzer. Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. Energy 2022;240:122500.
[13] Wang Yi, Chen Qixin, Hong Tao, Kang Chongqing. Review of smart meter data analytics: Applications, methodologies, and challenges. IEEE Trans Smart Grid 2018;10(3):3125–48.
[14] Li Ke, Mu Yuchen, Yang Fan, Wang Haiyang, Yan Yi, Zhang Chenghui. A novel short-term multi-energy load forecasting method for integrated energy system based on feature separation-fusion technology and improved CNN. Appl Energy 2023;351:121823.

[15] Qin Jiaqi, Zhang Yi, Fan Shixiong, Hu Xiaonan, Huang Yongqiang, Lu Zexin, Liu Yan. Multi-task short-term reactive and active load forecasting method based on attention-LSTM model. Int J Electr Power Energy Syst 2022;135:107517.
[16] Wan Anping, Chang Qing, Khalil AL-Bukhaiti, He Jiabo. Short-term power load forecasting for combined heat and power using CNN-LSTM enhanced by attention mechanism. Energy 2023;282:128274.
[17] Zhang Xinan, Chau Tat Kei, Chow Yau Hing, Fernando Tyrone, Iu Herbert Ho-Ching. A novel sequence to sequence data modelling based CNN-LSTM algorithm for three years ahead monthly peak load forecasting. IEEE Trans Power Syst 2023;39(1):1932–47.
[18] Molina-Markham Andrés, Shenoy Prashant, Fu Kevin, Cecchet Emmanuel, Irwin David. Private memoirs of a smart meter. In: Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building. 2010, p. 61–6.
[19] Husnoo Muhammad Akbar, Anwar Adnan, Hosseinzadeh Nasser, Islam Shama Naz, Mahmood Abdun Naser, Doss Robin. A secure federated learning framework for residential short term load forecasting. IEEE Trans Smart Grid 2023.
[20] Janghyun K, Barry H, Tianzhen H, et al. A review of preserving privacy in data collected from buildings with differential privacy. J Build Eng 2022;56:104724.
[21] McMahan Brendan, Moore Eider, Ramage Daniel, Hampson Seth, y Arcas Blaise Aguera. Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. PMLR; 2017, p. 1273–82.
[22] Chen Dayin, Shi Xiaodan, Zhang Haoran, Song Xuan, Zhang Dongxiao, Chen Yuntian, Yan Jinyue. A phone-based distributed ambient temperature measurement system with an efficient label-free automated training strategy. IEEE Trans Mob Comput 2024.
[23] Fekri Mohammad Navid, Grolinger Katarina, Mir Syed. Distributed load forecasting using smart meter data: Federated learning with recurrent neural networks. Int J Electr Power Energy Syst 2022;137:107669.
[24] Fernández Joaquín Delgado, Menci Sergio Potenciano, Lee Chul Min, Rieger Alexander, Fridgen Gilbert. Privacy-preserving federated learning for residential short-term load forecasting. Appl Energy 2022;326:119915.
[25] Wang Yi, Gan Dahua, Sun Mingyang, Zhang Ning, Lu Zongxiang, Kang Chongqing. Probabilistic individual load forecasting using pinball loss guided LSTM. Appl Energy 2019;235:10–20.
[26] Shen Tao, Zhang Jie, Jia Xinkang, Zhang Fengda, Huang Gang, Zhou Pan, Kuang Kun, Wu Fei, Wu Chao. Federated mutual learning. 2020, arXiv preprint arXiv:2006.16765.
[27] Raza Ali, Jingzhao Li, Ghadi Yazeed, Adnan Muhammad, Ali Mansoor. Smart home energy management systems: Research challenges and survey. Alex Eng J 2024;92:117–70.
[28] Chandrasekaran Radhika, Paramasivan Senthil Kumar. Advances in deep learning techniques for short-term energy load forecasting applications: A review. Arch Comput Methods Eng 2024;1–30.
[29] Han Dae-Man, Lim Jae-Hyun. Smart home energy management system using IEEE 802.15. 4 and zigbee. IEEE Trans Consum Electron 2010;56(3):1403–10.
[30] Zhou Bin, Li Wentao, Chan Ka Wing, Cao Yijia, Kuang Yonghong, Liu Xi, Wang Xiong. Smart home energy management systems: Concept, configurations, and scheduling strategies. Renew Sustain Energy Rev 2016;61:30–40.
[31] Ali Abdelrahman O, Elmarghany Mohamed R, Abdelsalam Mohamed M, Sabry Mohamed Nabil, Hamed Ahmed M. Closed-loop home energy management system with renewable energy sources in a smart grid: A comprehensive review. J Energy Storage 2022;50:104609.
[32] Bao Guannan, Lu Chao, Yuan Zhichang, Lu Zhigang. Battery energy storage system load shifting control based on real time load forecast and dynamic programming. In: 2012 IEEE international conference on automation science and engineering. CASE, IEEE; 2012, p. 815–20.
[33] Dai Shuang, Meng Fanlin, Wang Qian, Chen Xizhong. Federatednilm: A distributed and privacy-preserving framework for non-intrusive load monitoring based on federated deep learning. In: 2023 international joint conference on neural networks. IJCNN, IEEE; 2023, p. 01–8.
[34] Beckel Christian, Sadamori Leyna, Staake Thorsten, Santini Silvia. Revealing household characteristics from smart meter data. Energy 2014;78:397–410.
[35] Khan Noman, Haq Ijaz Ul, Khan Samee Ullah, Rho Seungmin, Lee Mi Young, Baik Sung Wook. DB-Net: A novel dilated CNN based multi-step forecasting model for power consumption in integrated local energy systems. Int J Electr Power Energy Syst 2021;133:107023.
[36] Taïk Afaf, Cherkaoui Soumaya. Electrical load forecasting using edge computing and federated learning. In: ICC 2020 - 2020 IEEE international conference on communications. ICC, 2020, p. 1–6. http://dx.doi.org/10.1109/ICC40277.2020.9148937.
[37] Mansour Yishay, Mohri Mehryar, Ro Jae, Suresh Ananda Theertha. Three approaches for personalization with applications to federated learning. 2020, arXiv preprint arXiv:2002.10619.
[38] Wang Yi, Gao Ning, Hug Gabriela. Personalized federated learning for individual consumer load forecasting. CSEE J Power Energy Syst 2022.
[39] Qu Xiaodong, Guan Chengcheng, Xie Gang, Tian Zhiyi, Sood Keshav, Sun Chaoli, Cui Lei. Personalized federated learning for heterogeneous residential load forecasting. Big Data Min Anal 2023;6(4):421–32.
[40] Tang Lingfeng, Xie Haipeng, Wang Xiaoyang, Bie Zhaohong. Privacy-preserving knowledge sharing for few-shot building energy prediction: A federated learning approach. Appl Energy 2023;337:120860.
[41] Qin Dalin, Wang Chenxi, Wen Qingsong, Chen Weiqi, Sun Liang, Wang Yi. Personalized federated darts for electricity load forecasting of individual buildings. IEEE Trans Smart Grid 2023;14(6):4888–901.
[42] Gou Jianping, Yu Baosheng, Maybank Stephen J, Tao Dacheng. Knowledge distillation: A survey. Int J Comput Vis 2021;129:1789–819.
[43] Cho Kyunghyun, Van Merriënboer Bart, Gulcehre Caglar, Bahdanau Dzmitry, Bougares Fethi, Schwenk Holger, Bengio Yoshua. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014, arXiv preprint arXiv:1406.1078.
[44] Hinton Geoffrey, Vinyals Oriol, Dean Jeff. Distilling the knowledge in a neural network. 2015, arXiv preprint arXiv:1503.02531.
[45] Zhang Yu, Tang Guoming, Huang Qianyi, Wang Yi, Wu Kui, Yu Keping, Shao Xun. Fednilm: Applying federated learning to nilm applications at the edge. IEEE Trans Green Commun Netw 2022.
[46] Murray David, Stankovic Lina, Stankovic Vladimir. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. Sci Data 2017;4(1):1–12.