Fed-SAD:A secure aggregation federated learning method for distributed load forecasting

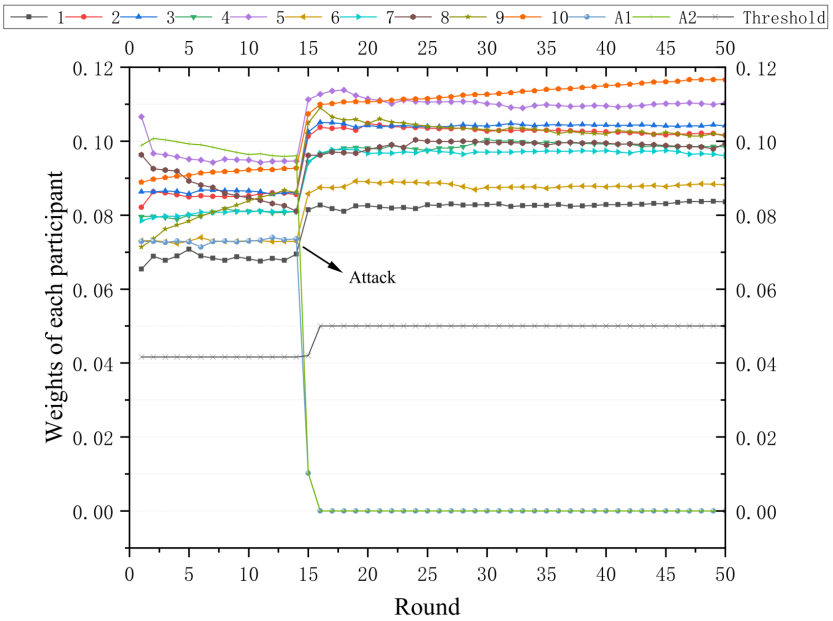
Jian Li[1], Hexiao Li[1], Ruiqi Wang[2], Yiguo Guo[2], and Sixing Wu[1]

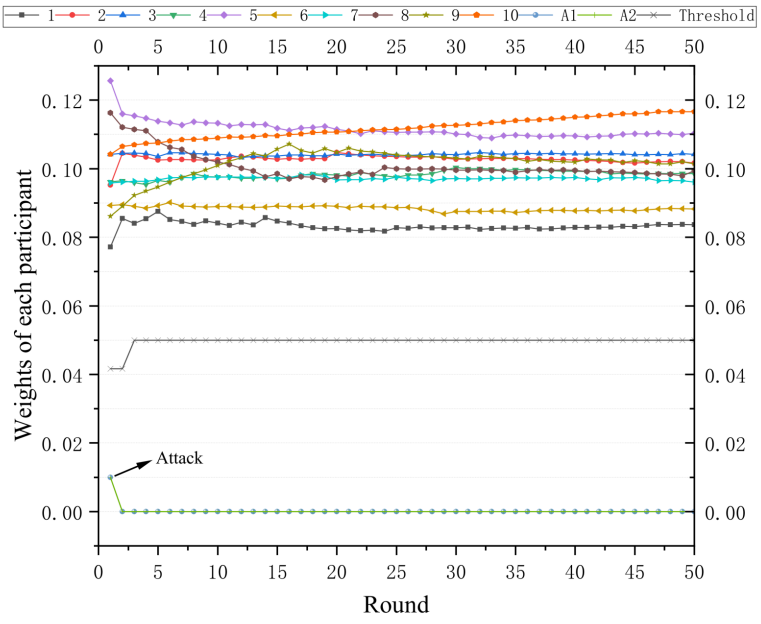
[1]North China Electric Power University
[2]State Grid Shandong Electric Power Company

July 25, 2023

Abstract

The distributed and privacy-preserving characteristics of fine-grained smart grid data hinder data sharing, making federated learning an attractive approach for collaborative training among data owners with similar load patterns. However, malicious models can interfere with training in the federated learning aggregation process, making it difficult to ensure the accuracy and safety of the central model in load forecasting. Therefore, we propose a secure aggregation federated learning method for distributed load forecasting based on similarity and distance (Fed-SAD), which effectively eliminates the interference of malicious models by securely aggregating models, thereby ensuring accurate and safe distributed scenario prediction. Experimental results demonstrate that Fed-SAD maintains high accuracy and robustness in both the presence and absence of malicious models, while maintaining data and model security.





IET Research Journals

Brief Paper

IET Journals
The Institution of Engineering and Technology

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Fed-SAD:A secure aggregation federated learning method for distributed load forecasting

Jianbin Li[1], *Hexiao Li*[1], *Ruiqi Wang*[2], *Yiguo Guo*[2], *Sixing Wu* [1*]

[1] *School of Control and Computer Engineering, North China Electric Power University, Beijing, China*
[2] *Economic Technology Research Institute, State Grid Shandong Electric Power Company, Jinan city, Shandong Province, China*
* *E-mail: wusx@ncepu.edu.cn*

Abstract: The distributed and privacy-preserving characteristics of fine-grained smart grid data hinder data sharing, making federated learning an attractive approach for collaborative training among data owners with similar load patterns. However, malicious models can interfere with training in the federated learning aggregation process, making it difficult to ensure the accuracy and safety of the central model in load forecasting. Therefore, we propose a secure aggregation federated learning method for distributed load forecasting based on similarity and distance (Fed-SAD), which effectively eliminates the interference of malicious models by securely aggregating models, thereby ensuring accurate and safe distributed scenario prediction. Experimental results demonstrate that Fed-SAD maintains high accuracy and robustness in both the presence and absence of malicious models, while maintaining data and model security.
Keywords: Data Privacy, Federated Learning, Distributed Load Forecasting, Model Security Aggregation , Malicious Model

1 Introduction

As the power system continues to develop, data analysis and security have become increasingly important issues. Accurate and safe load demand forecasting using historical data in the power system can help power companies and operators make rational operational decisions and optimal plans[26]. Researchers have applied various algorithms like BP neural network and decision trees to the field of power data analysis and improved them to be more adaptable, enabling the use of artificial intelligence methods for accurate data prediction and analysis[1]. With the fine-grained segmentation and distributed management of modern power grids, data exhibits a distributed nature. For instance, data samples obtained from various regions, household meters, and substations are subject to high data confidentiality requirements, which limit data sharing and restrict the utility of data. To address this challenge, federated learning - an approach proposed by Google in 2016 - presents a promising solution that effectively balances the conflict between power data availability and data privacy. The participants can jointly train models by aggregating model parameters trained on local data, enabling them to share training parameters and expand data samples[21]. To leverage the benefits of federated learning and meet the demands of distributed forecasting scenarios, current research applies federated learning to forecasting datasets with similar energy consumption characteristics[6]. Such an approach can effectively resolve privacy issues in load forecasting scenarios and enhance predictive capabilities[23].

Despite its potential benefits, federated learning faces attackers uploading malicious models in the model aggregation process. Such issues could impact the accuracy of the overall model. During the model training process, if the central server aggregates the uploaded parameters without evaluating their reasonableness, it could affect the accuracy of the overall model parameters. Subsequently, other participants who use the centralized model parameters issued by the server for local model training could experience reduced model accuracy[16]. The current aggregation methods for load prediction models in federated learning have certain limitations. For instance, the classical Federated Average Aggregation Algorithm (FedAvg) weights the model based solely on the sample size of the training data volume of each participant. However, this approach does not take into account practical scenarios and does not consider the quality of the data samples provided by the participants. Furthermore, it does not address issues such as the presence of malicious models that might interfere with the overall model aggregation quality in real-world applications. Such limitations could ultimately reduce the predictive and generalization capabilities of the final model. Prior research has revealed that using the FedAvg algorithm to directly aggregate the parameters of local models can lead to a loss of training performance for the entire federated learning process if some participant nodes upload malicious model parameters, which would negatively impact the training results of other participants[17]. There are a limited number of research methods that address secure aggregation based on federated learning in the context of distributed load forecasting, and these methods have certain limitations. For instance, the method that spectral clustering algorithm to detect poisoning attacks fails to fully account for the quality of each participant's parameters during the aggregation process[18], while FedClamp relies on a testing and validation method to ensure the reasonableness of uploaded parameters, which lengthens each round[14].

This paper proposes a federated learning secure aggregation algorithm to tackle the issue of malicious models in load prediction collaboration based on Federated Learning. At each round of central model aggregation, the proposed algorithm first calculates the parameter similarity between two participants using the parameters uploaded by local participants. It then analyzes problem by solving the "maximum clique problem" using graph theory. The model parameters in the "maximum clique" are then used to approximate the parameters of the global model for the current training round. To achieve this, the algorithm utilizes cosine similarity as the metric for model features, which has been shown to be a better choice[19]. Cosine similarity takes into account the direction and is not affected by scaling effects. Mathematically, the cosine similarity measures the angle between two input vectors, and the smaller the angle, the higher the cosine similarity value. Therefore, using cosine similarity to calculate the approximate global model parameters leads to relatively accurate results and accelerates the model's convergence. Furthermore, the algorithm employs the Gaussian probability distribution function to assign a larger weight to participants with smaller distances between the approximate global

model calculated in each round and their respective local models. The proposed method further improves the quality of the central model.

Based on the aforementioned analysis, applying federated learning directly for model training and aggregation in the current data collaborative training scenarios in the electric power industry presents potential vulnerabilities. Such an approach could enable malicious participants to upload malicious models and interfere with the entire federated learning training process. Therefore, this paper proposes a secure aggregation federated learning method for distributed load forecasting based on similarity and distance (Fed-SAD) that ensures model security aggregation while maintaining the privacy of load data owners.

To summarize, this paper makes the following three contributions:

(1) We propose Fed-SAD, a secure aggregation method for federated learning. The global model parameters of the approximate solution is used as the criteria for selecting participants in each round, which facilitates the prompt identification of malicious models and accelerates model convergence.

(2) A Gaussian probability density function is introduced into the method, which assign different weights to each participant in order to further improve the quality of the model.

(3) We conduct experiments in collaborative training scenarios for load forecasting in federated learning to prove the effectiveness and robustness of the propsed method. Fed-SAD effectively excludes malicious models while enhancing prediction accuracy.

2 Related work

2.1 Secure aggregation method research based on federated learning

In the context of federated learning, the training process is designed to prevent the central server from accessing data from participants directly, limiting it to only aggregating model parameters from participants. However, the inability to detect malicious models during the parameter aggregation process could negatively impact the entire training process, resulting in subpar model quality and reduced training accuracy for all participants. Currently, security aggregation algorithms for federated learning mainly include the aggregation algorithm based on the model update feature difference and the aggregation algorithm based on the verification data set[9].

The model update feature difference aggregation algorithm leverages the differences between normal and malicious model updates to allow the central server to distinguish the latter. One example of this type of algorithm is the multi-Krum model, which calculates the Euclidean distance between all local models and excludes edge nodes that are relatively far from the overall distribution[3]. Meanwhile, the Cao scheme treats participants as nodes in a graph. When the Euclidean distance between two participants' model weight parameters is less than a defined threshold, a link is added between them. The server subsequently identifies the largest clique on the generated graph, then takes the average of the node parameters on that clique as the aggregation result[5] . Another approach is FoolsGold, which first calculates historical aggregated updates for each participant in each iteration before adjusting weights based on maximum cosine similarity with other participants. Less similarity implies a greater likelihood of malicious updates and lower weights, with the final result being a weighted average of the updates[7]. Median aggregates its input gradients by computing the median of the values of each dimension of the gradient[25]. However, Baruch M has proven that the above methods have limitations: historical aggregated updates can hinder faster model convergence[2] .

The aggregation algorithm based on the validation dataset requires the central server to possess data samples that are similar to those owned by the participants. This algorithm utilizes a validation dataset to evaluate parameters uploaded by the participants and identify erroneous updates. One shared model detection technique based on feature importance is proposed by Fed-Fi. It can effectively mitigate model accuracy degradation due to conspiracy attacks, provided the server has a small amount of sample data[27]. Furthermore, Su proposes setting up a high-quality test dataset in the central server to determine the reputation value of each participant, which will then be used to lower the weight of participants with lower reputation values in the global model aggregation process[20]. However, these methods may not apply to the power system scenario where the central server lacks access to the actual energy consumption data.

2.2 Load forecasting-oriented secure aggregation method research based on federated learning

Some studies have addressed the issue of detecting malicious models and performing model-safe aggregation during load forecasting based on Federated Learning. For instance, N. B. S. Qureshi suggests that in a Federated Learning collaborative training load forecasting scenario, the central server clusters the uploaded parameter weight features into two sets and identifies nodes in the smaller set as anomalous participants before aggregating the parameters in the larger set directly with FedAvg[18]. Meanwhile, FedClamp proposes using Hidden Markov Models for anomalous model detection to identify anomalous participants before model aggregation[14]. However, these approaches only perform initial screening of potentially malicious models, and the aggregation weights do not fully consider the impact of each participant on the weight of central model. As a result, the aggregation process may waste part of the data sample features for a small number of participants who are not malicious models.

In the domain of federated learning, both in general scenarios and specifically for load prediction, research has been conducted on secure aggregation methods. These studies primarily focus on designing secure aggregation algorithms to safeguard the security of models. In our research, we aim to enhance existing methodologies and introduce a novel federated learning-based secure aggregation approach tailored for load prediction. In comparison to current techniques, our proposed approach utilizes a similarity and distance-based aggregation algorithm. Our proposed approach enhances model accuracy, and effectively mitigates the influence of malicious models on the central model. Through experimentation, our method exhibits notable advantages in load forecasting tasks and offers a practical solution to achieve secure and efficient load prediction.

3 Proposed secure federated aggregation method for distributed load forecasting

3.1 Overview of proposed method architecture and design

The secure aggregation method proposed in this paper ensures that each participant does not have to share their local data, thereby preserving the privacy of sensitive information. It guarantees the accuracy and validity of the central model by eliminating the interference of malicious models and expanding the sample data space of each load data owner, leading to improved prediction accuracy. Importantly, our method uploads the complete parameter weights of the model rather than gradients to protect against potential gradient parameter leakage. Such a breach would enable attackers to infer corresponding training data and compromise data security[13].

The Fig.1 illustrates the overall architecture of our method, where participants can consist of various users, such as electric utilities, distributed buildings, and communities. The aggregation of model parameters can better predict load patterns for each participant and enhance prediction performance when the data characteristics among participants are more similar. Each participant uses their own local data directly to train their local model.

The overall process consists of four main steps, as shown in Algorithm 1:

(1) The LSTM algorithm is used as the foundation for model training. The central server initializes the model parameters and distributes them to each participant node.

(2) Participants receive the model parameters from the central server, update their local parameters with this information, and train the model using their local data.

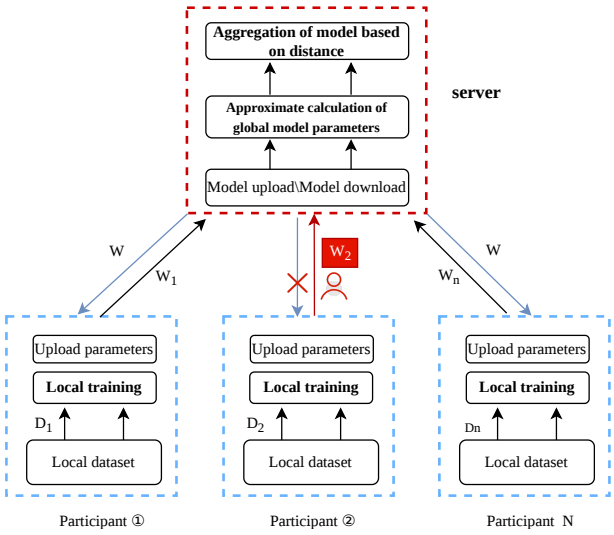


Fig. 1: Architecture of the Fed-SAD

(3) After each round of training, participants upload their updated parameters to the central server.

(4) The central server approximates the global model parameters based on the parameters uploaded by the participants. Following that, it assigns weights to the parameters of individual participants using the Euclidean distance between the approximate global model parameters and the local model parameters. Finally, it calculates the global model parameters based on this weighted aggregation.

Algorithm 1 Overview of the proposed federated learning security aggregation algorithm

Input: The number of rounds of execution T, The distributed load data participants involved in the training L, Learning Rate η .
Output: The model trained by the method proposed in this paper.

```
Server:
The central server initializes the model;
Set the initial local participants involved in the training p_i(k = 1, 2, ··· , n);
Receive model parameters from participants;
Filtering and aggregation of models using Algorithm 2;
Participants:
for t = 1, 2, 3 ··· , T do
    for p_i ∈ L do
        p_i Download and update local weights from server w_t;
        p_i Update and train local models and upload parameters to central server.
    end for
end for
```

3.2 *Local training*

For local model training, we utilize the LSTM (Long Short-Term Memory) model[8], which is a type of recurrent neural network (RNN) that excels at time series prediction tasks. LSTM is an improved form of the recurrent neural network, which introduces gated self-circulation to ensure that gradients can be propagated for a long period and solve the problem of gradient vanishing. Therefore, LSTM models can better extract long-term dependency features in learning sequences and are widely used in time series prediction problems. Similarly, in distributed load forecasting, this neural network can be used to learn how daily load consumption patterns can impact future energy usage. Therefore, we have chosen this model algorithm for local model training in our research.

Fig. 2 shows the internal structure of the constituent units of the LSTM model. The cells of LSTM are connected to each other, which

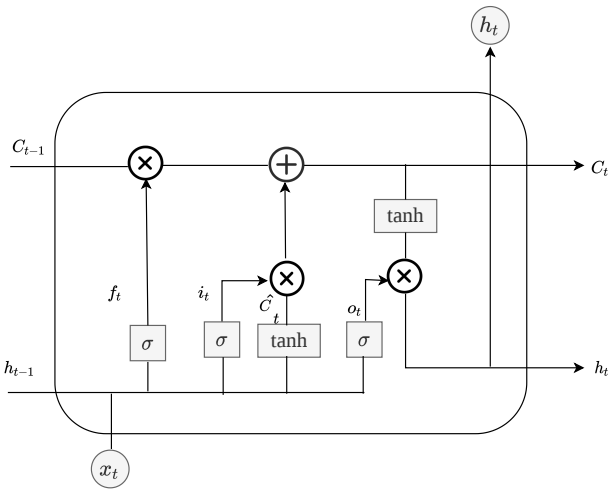


Fig. 2: LSTM network basic unit

mainly consist of input gates, forgetting gates and output gates.σ is the sigmoid activation function, x_t is the input sequence value at that moment, and tanh represents the tanh activation function. C_{t-1} and C_t represent the cell states at the previous and current moments. In addition, h_{t-1} and h_t denote the hidden states at the previous and current moments. f_t, i_t and o_t denote the forgetting gate, input gate and output gate calculation variables at the current moment, and $\hat{C}_t$ denotes the input cell state at that moment. The above variables are calculated as follows:

$$f_t = \sigma(W_f h_{t-1} + W_f x_t + b_f) \tag{1}$$

$$i_t = \sigma(W_i h_{t-1} + W_i x_t + b_i) \tag{2}$$

$$o_t = \sigma(W_o h_{t-1} + W_o x_t + b_o) \tag{3}$$

$$\hat{C}_t = tanh(W_c h_{t-1} + W_c x_t + b_c) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

In equation (1-6),W_f, W_i, W_o denote the weight matrices of forget gate, input gate, and output gate.b_f, b_i, b_o denote the output bias vectors of the forget gate, input gate and output gate. The $*$ denotes the Hadamard product of matrices.

In this paper, load forecasting refers to the use of historical load data to predict load values for future time periods. The input data in our study primarily consists of date-time features and historical load data. Since the load data and date-time features may differ greatly in magnitude, we normalize the input data to account for these differences and ensure that participants with varying magnitudes of load data can still contribute to the federated learning process.

3.3 *Securing federated learning with Fed-SAD aggregation algorithm*

In current research on secure aggregation algorithms for federated learning, some methods rely on either model parameter similarity or statistical characteristics to select model parameters for the current round. However, these approaches need to use the global model of the previous round as a benchmark, which results in a certain

lag in selecting model participants[24]. Alternatively, some methods perform average aggregation directly after selection without fully considering the quality of different model parameters. In contrast, our study proposes a secure aggregation federated learning approach for distributed load forecasting. In this approach, all local load data owners upload their models to the central server after completing local training, and the central server performs model screening and aggregation weighting on all received models. If there is any interference from malicious models in a particular round, we remove the model parameters of the participants associated with the malicious model before aggregation. The central model parameters of this round are weighted to aggregate the model parameters of participants. The weighting method is based on the offset between the model parameters uploaded by each participant and the computed approximate global model parameters. This ensures that the final aggregated central model has improved results, improves the training accuracy of each participant, and ensures secure model aggregation. The Fed-SAD model aggregation algorithm consists of two key steps, which are detailed below. Additionally, Algorithm 2 provides a specific algorithmic procedure for the approach.

Algorithm 2 Fed-SAD: A securing aggregation algorithm for federated learning in distributed load forecasting

Input: Model parameters uploaded by each participant w_t^i.
Output: Global training model w_t.

Initialization :$\beta = 1/2, Step = 0.05, Trust \Leftarrow \varnothing, V = \{p_1, p_2, p_3, \cdots p_n\}$
for $t = 1, 2, 3 \cdots, T$ **do**
  while $trust == \varnothing$ **do**
    for i *in* $range(N-1)$ **do**
      $G(p_i) \Leftarrow \varnothing$
      for j *in* $range(i+1, N)$ **do**
        if $Similarity(w_t^i, w_t^j) > threshold$ **then**
          $G(p_i) \Leftarrow G(p_i) \cup p_j$
        end if
      end for
    end for
    $Cliques \Leftarrow BronKerbosch(V, G)$
    $MaxClique \Leftarrow FindLargestClique(Cliques)$
    // Find the largest clique
    if $|MaxClique| >= N/2$ **then**
      $Trust = MaxClique$
      $\tilde{w_t} = \Sigma_{i=1}^{trust} \frac{1}{|trust|} w_t^i$
      Return $\tilde{w_t}$
      //Approximate the global model for each round
    end if
    $\beta = \beta + Step$
  end while
  for $i = 1, 2, 3 \cdots, N$ **do**
    $guassian(w_t^i; \tilde{w_t}, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\left\|w_t^i - \tilde{w_t}\right\|_2\right)^2}{2\sigma^2}}$
    $\alpha_t^i = softmax(guassian(w_t^i; \tilde{w_t}, \sigma))$
    if $\alpha_t^i < \beta$ **then**
      $N = N \backslash i$ // Remove malicious models
    end if
    $w_t = \sum_{i=1}^{N} \alpha_t^i w_t^i$ // Aggregate weights
  end for
end for

3.3.1 Approximate calculation of global model parameters for each round: To determine convergence trends as soon as possible and reduce the interference of malicious models, an approximate calculation of the global model parameters for each round is performed. This calculation is carried out on the parameters submitted by the participants to estimate the overall trend.

Our research employs a graph-theoretic knowledge modeling scheme to construct a graph where each vertex represents a local model uploaded by participants in the federated learning process. Each edge in the graph is generated based on the cosine similarity of the uploaded model parameters between two vertices. As indicated by Equation 7. To calculate the similarity between each pair of local models, we first set a threshold value. If the similarity value exceeds the threshold β between two models, an edge is created between the two points; otherwise, no edge is created.

$$Similarity(w_t^i, w_t^j) = cos(w_t^i, w_t^j) = \frac{\left\langle w_t^i, w_t^j \right\rangle}{\left|w_t^i\right| \left|w_t^j\right|} \tag{7}$$

To find a matching set of cliques, we solve the maximum clique problem by dividing the local models into different cliques. We use BronKerbosch[4] algorithm to find all the cliques and then find the clique that contains the most local models while meeting the condition that the number of vertices in this set is greater than N/2. If the number of vertices in this set is less than N/2, decrease β with Step,we decrease the similarity threshold and iterate until we find a matching of cliques.

Ultimately, we obtain an approximation of the global model parameters for this round by finding the largest clique. This approach helps us identify a reliable set of local model parameters and improve the estimation of global model parameters.

3.3.2 Aggregation of models based on distance : In our research, we use similarity to calculate the approximate global model parameters for each round instead of directly aggregating the global model. This is because if the model parameters uploaded by individual participants significantly differ from the global model, the global model can move in the wrong direction from the desired model convergence. Using the wrong global model convergence direction as the criterion for anomaly model screening is undesirable.

In real-life federated learning scenarios, research statistics show that the percentage of malicious models does not usually exceed $1/5$ [10]. Therefore, it is appropriate and accurate to use the "maximum clique problem" to calculate the approximate global model parameters. By dividing local models into different cliques and identifying the largest clique, we obtain a reliable set of local model parameters that can help improve the estimation of global model parameters while minimizing the interference from faulty or malicious models.

Subsequently, to measure the anomalous nature of the model, our research considers the use of a Gaussian probability distribution function. We calculate the distance between each participant's local model and the approximate global model for the current round. Based on this distance, we assign different weights to the participants, with higher weights for those whose distance from the central model parameters is smaller. In this way, we exclude model parameters that deviate far from the central model parameters and obtain a reliable aggregation result. The final aggregation result is a weighted average of the uploaded parameters of each participant. By using a probability density function based on the Gaussian distribution, we can identify anomalies in the model parameters and assign appropriate weights to each participant's input to obtain more accurate and reliable global model parameters.

In comparison to normal participants, anomaly models exhibit greater differences in their parameters from the central server. As a result, our research measures anomalous models based on the distance between the approximate global model and the local model calculated in the previous subsection. We then perform the aggregation of weights. During each training round's aggregation process, we assign smaller weights to participants whose local models have larger distances from the global model. This is because the greater the distance between the global and local models, the higher the degree of anomaly. By utilizing this method, we can identify and exclude anomalous models from the aggregation process, thereby improving the accuracy and reliability of the global model parameters. The process is illustrated in Fig.3. The central server first performs each round of receiving parameters. Secondly calculates the similarity of the parameters of every two participants using the parameters uploaded by the local participants, uses graph theory to solve the "maximum clique problem " to find the maximum clique, and uses the parameters of the participant models in this cluster

set to calculate the parameters of the global model that approximates the current training round. Thirdly uses the distance difference between the parameters of the global model from the approximate solution and the local model, for malicious models that exceed the threshold to eliminate them to avoid the impact on the central model aggregation.

In this paper, we utilize the Gaussian distribution function to establish a correlation between the distance and the degree of model anomaly. Equation 8 defines the expression for the Gaussian distribution function.

$$guassian(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

The approximate global model parameter for this round is $\tilde{w}_t$, while the parameter uploaded by each participant for this round is w_i^t. Equation 9 calculates the Euclidean distance between all clients and the approximate global model parameters.

$$d = \left\| w_t^i - \tilde{w}_t \right\|_2 \tag{9}$$

Substituting this value into the equation for the Gaussian distribution function, we obtain Equation 10.

$$guassian(w_t^i;\tilde{w}_t,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\left\| w_t^i - \tilde{w}_t \right\|_2\right)^2}{2\sigma^2}} \tag{10}$$

The standard deviation of the Gaussian distribution function is computed using Equation 11.

$$\sigma = \sqrt{\frac{1}{N}\Sigma_{i=1}^{n}\left(\left\| w_t^i - \tilde{w}_t \right\|_2\right)^2} \tag{11}$$

The output of the Gaussian distribution function is then normalized to derive the participation weight of each participant, as shown in Equation 12. This final weight determines the degree of influence of each participant in the federated learning process.

$$\alpha_t^i = softmax(guassian(w_t^i;\tilde{w}_t,\sigma)) \tag{12}$$

4 Experiment

4.1 Dataset

In this paper, we conduct experiments using the open-source Building Data Genome Project dataset[15]. This dataset consists of energy consumption data from buildings in various countries. To address the challenge of varying energy consumption patterns, we select datasets with similar profiles to facilitate model training and improve prediction accuracy[11]. Specifically, we use 12 datasets with similar energy usage patterns[12], namely OfficeMax, OfficeMarcus, OfficeMonty,OfficeMaya,OfficeMyron, OfficeMick, OfficeMarion, OfficeMartha, OfficeMalik, OfficeMadaDetailed, OfficeMuhammad, andnOfficeMoses. Our experimentation included ten trusted participants and two adversaries. The data for these twelve buildings are recorded as $l_1, l_2, \cdots, l_{10}, A_1, A_2$. It is important to note that our proposed method aims to demonstrate that the accuracy of load prediction is not significantly impacted by the presence of malicious models, and that the aggregated model still provides good prediction results even in the absence of malicious models. Therefore, we excluded differences in the amount of data from each region and interference from selected dates, opting instead to use load data from the same period for this study. Ultimately, we selected the load dataset for July from these twelve buildings, using the first thirty days for training and the last day for testing.

4.2 Baseline and attack model

To evaluate the effectiveness of the proposed security aggregation algorithm based on federated learning, we conduct three scenarios, namely: (1) aggregating models without malicious model interference, (2) Sign Flipping Attack, and (3) Additive Noise Attack.

Table 1 Experimental parameters

Parameters	Setup/Values
Optimizer	Adaptive Moment Estimation (Adam)
Learning rate	0.06
communication round	50
local training epochs	6
$Threshold$	$1/2N$
β	1/2
$Step$	0.05
Batch of size	256

4.2.1 Baseline: To provide a basis for comparison, we introduce the following methods:

(1) Standalone: Each participant was trained independently and locally using only their own local dataset without collaboration.

(2) FedAvg (Federated Averaging Algorithm): We compare Fed-SAD with this method under both the presence and absence of malicious model attacks.

(3) FoolsGold: In this method, historical aggregated updates of each participant are calculated in each iteration, and then the weight of each participant is adjusted based on the maximum cosine similarity of its historical aggregated updates with others. We compare this method with Fed-SAD in the presence of malicious model attacks.

(4) Median: This method sorts the jth parameter of each local model and uses the sorted median as the jth parameter of the global model. For comparison with Fed-SAD in the presence of malicious model attacks.

4.2.2 Attack Models Setup: In the attack models setup, we consider two attack scenarios:

(1) Sign Flipping Attack: In the case of simulating this attack, the anomaly participant inverts the sign of the parameters in the local model. The aim is to make the parameters transform in the opposite direction, changing the convergence direction of the global model and destroying the training efficiency of federal learning and the accuracy of the model. The sign flipping attack can be mathematically represented by Equation 13.

$$w_t^{i'} = -1 * w_t^i \tag{13}$$

(2) Additive noise attack: The anomaly model influences the training of the global model by adding noise to the uploaded model parameters. Random numbers that obey Gaussian distribution and have the same size as the training model are added to the weight values as parameters of the attack model. The mathematical expression is shown in Equation 7.

$$w_t^{i'} = w_t^i + Gaussian(0, NoiseIntensity) \tag{14}$$

4.3 Training setting and evaluation criteria

The experimental setup of this study was conducted on a Windows 11-64 bit operating system, using an 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz 2.30 GHz processor and a NVIDIA GeForce RTX 3050 Ti Laptop GPU. The deep learning algorithm was implemented in Python 3.8 programming language, with the PyTorch 1.13 deep learning framework. The parameters used for the experimental setup are listed in Table 1.

To evaluate the load forecasting effect, we use the mean absolute percentage error (MAPE) as the evaluation index. A smaller MAPE value indicates a smaller error between the predicted value and actual value, and thus better model predictions[22]. Equation 15 shows the mathematical formula that is used to calculate the MAPE. In the equation y_i^{pre} represents the forecasted load value and y_i^{real} represents the actual load value.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i^{real} - y_i^{pre}}{y_i^{real}}\right| \tag{15}$$

Table 2 Forecast accuracy for each participant without malicious models

MAPE/%	1	2	3	4	5	6	7	8	9	10
Standalone	8.14%	8.41%	8.45%	8.43%	7.01%	7.26%	8.96%	9.56%	8.58%	8.39%
FedAvg	7.98%	8.22%	8.09%	7.59%	5.90%	8.65%	8.93%	8.36%	8.49%	6.04%
Fed-SAD	6.87%	5.18%	4.06%	4.73%	3.56%	7.21%	4.36%	6.26%	5.03%	3.24%

Table 3 Forecast accuracy for each participant under Sign Flipping Attack with 10 honest participants (adversaries omitted)

MAPE/%	1	2	3	4	5	6	7	8	9	10
FedAvg	21.89%	19.48%	20.82%	24.86%	21.34%	22.50%	21.06%	23.12%	22.14%	20.90%
Median	15.47%	12.17%	14.37%	17.36%	14.71%	13.55%	14.35%	16.86%	14.66%	15.18%
FoolsGold	12.19%	10.69%	11.54%	12.37%	11.23%	12.79%	13.07%	14.12%	13.25%	14.60%
Fed-SAD	7.61%	5.04%	6.15%	4.98%	4.26%	6.74%	7.32%	5.37%	6.67%	4.05%

Table 4 Forecast accuracy for each participant under Additive Noise Attack with 10 honest participants (adversaries omitted)

MAPE/%	1	2	3	4	5	6	7	8	9	10
FedAvg	15.13%	17.18%	16.79%	18.23%	18.77%	15.59%	17.49%	19.06%	18.45%	16.78%
Median	12.15%	11.50%	12.55%	16.16%	14.31%	13.79%	15.69%	12.46%	13.87%	14.39%
FoolsGold	10.97%	11.64%	9.87%	8.93%	10.47%	12.47%	9.81%	8.76%	11.79%	13.07%
Fed-SAD	7.06%	5.24%	4.19%	4.56%	3.87%	7.41%	5.19%	6.35%	5.12%	3.31%

4.4 Analysis of results

This paper evaluates the proposed security aggregation method, Fed-SAD, in two aspects. Firstly, we evaluate and compare the prediction accuracy of Fed-SAD in the absence of attack model interference, as well as under Additive noise attack and Sign Flipping Attack. Secondly, we evaluate the robustness of the Fed-SAD method in quickly detecting and eliminating participants who upload malicious models.

4.4.1 Predictive performance: To evaluate the predictive performance, we compare and analyze the accuracy of model load prediction in the following three cases.

(1) No Attack Model: Firstly, we train and predict the model without any attack interference in this paper and compare the Fed-SAD method with FedAvg and Standalone methods. Table 2 shows the prediction results. The evaluation index used in this paper is MAPE. It can be observed from the results that the Fed-SAD method improves the prediction accuracy by up to 5% over the Standalone method and up to 4.5% over the FedAvg method.

(2) Impact by Sign Flipping Attack: For the sign-flipping attack, the experiments involve twelve participants, with two adversaries performing the attack after each round of training. Among the ten trusted participants, the Fed-SAD method shows better performance compared to FedAvg, FoolsGold, and median methods, with prediction results 13%-19%, 4%-10%, and 6%-12% higher, respectively. The effectiveness of the approach was demonstrated in the results presented in Table 3.

(3) Impact by Additive noise attack Similarly, for the Gaussian noise attack simulation, twelve participants are employed, and two adversaries perform the attack after each training round. During the experiments, we set the variance value to 0.1 to represent additive noise. The prediction results of the Fed-SAD method show better performance than the FedAvg, FoolsGold, and median methods, with predictions 8%-15%, 2%-10%, and 5%-11% higher, respectively. The results are summarized in Table 4. Overall, the study shows that the Fed-SAD approach is effective in improving model load prediction accuracy in the presence of adversarial attacks.

From the analysis of the prediction accuracy results presented in the study, it can be observed that Fed-SAD perform better in the presence of adversarial attacks as compared to other methods such as FedAvg, Median, and FoolsGold. FedAvg is the most affected by malicious models since it aggregates weights directly without a defense strategy. The Median strategy is better than FedAvg in reducing the interference of malicious models to some extent, but it cannot achieve the optimal aggregation of aggregated model parameters and cannot obtain the best prediction effect. On the other hand, while FoolsGold performs better than Median and FedAvg, the Fed-SAD method perform slightly better than FoolsGold in terms of prediction accuracy. In distributed load forecasting, we consider scenarios with similar data characteristics and use distance as the basis for the final aggregation weight in the Fed-SAD method.

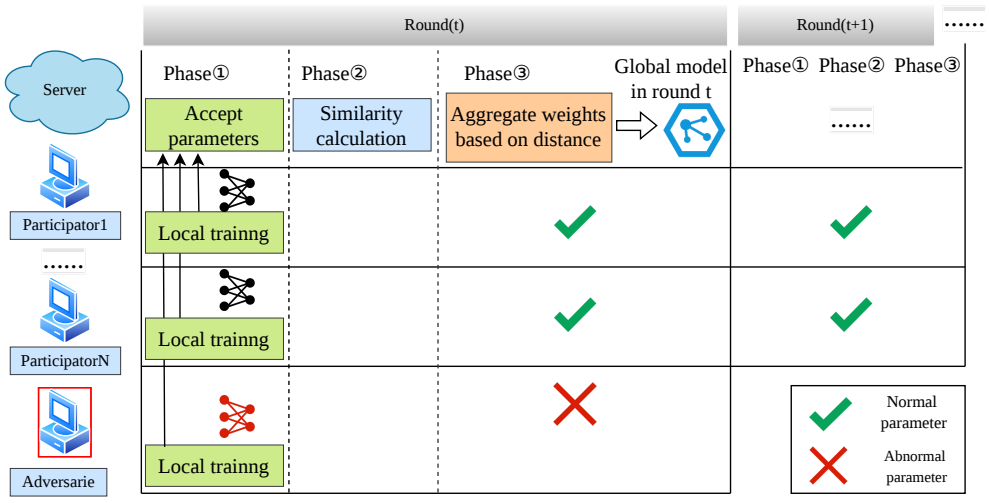


Fig. 3: Workflow for removing attack models

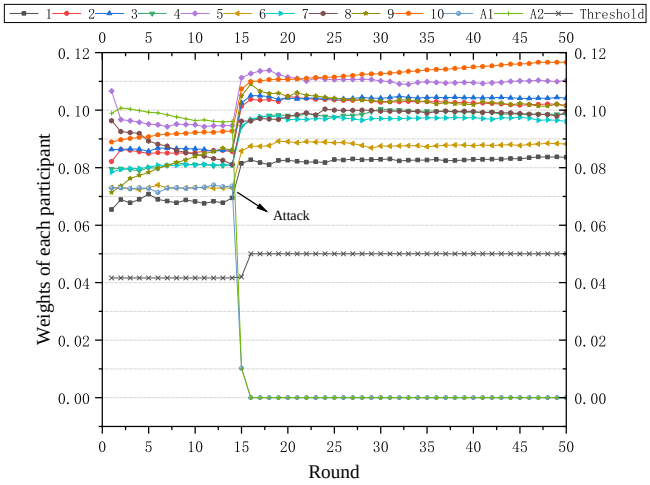


Fig. 4: Weight changes of each participant in the Fed-SAD approach during federated learning model aggregation under sign flipping attacks

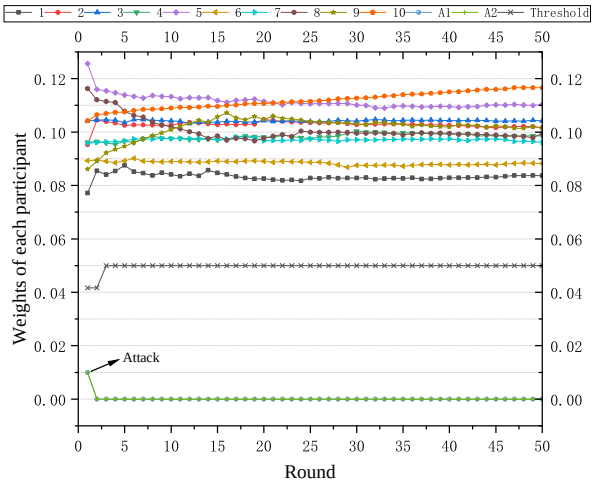


Fig. 5: Weight changes of each participant in the Fed-SAD approach during federated learning model aggregation under Additive noise attacks

This approach results in improved accuracy for the final prediction. Therefore, the results of the study demonstrate the effectiveness of the Fed-SAD approach in improving model load prediction accuracy in the presence of adversarial attacks.

It is worth noting that in the above prediction accuracy experiments, a relatively moderate attack intensity is set in this paper. To avoid wasting the diversity of data samples, the threshold is set low, and the attackers' weights are kept below the level of ten normal participants to participate in the training process with a lower threshold. For attack models with high attack intensity, the Fed-SAD method can directly remove them. To demonstrate the effectiveness of removing the anomalous models, the following experiments involving adjusting the malicious attack model settings were conducted to directly remove the malicious models.

4.4.2 Adversarial robustness: The robustness of Fed-SAD is analyzed in the study by varying the weight of different participants in the aggregation of a federated learning model, while simulating two types of attacks - additive noise attack and sign flipping attack. Fig.4 presents the variation of weight size for each participant when simulating the presence of sign flipping attacks. To reflect the effect of removing attack participants in the method, the weights are multiplied by a constant (set to 10) after sign-flipping. The experiments are conducted in the fifteenth round of model training, where two participants perform attacks on uploaded parameters that are identified and removed from collaborative training by the central server. Fig.5 shows the graph of weight size changes for each participant during the Additive noise attack simulation. In the experiment, two participants performed Additive noise attacks on the uploaded parameters at the beginning of the model training. To better observe the effect of weight change in the attacked model, the noise intensity was set to 20 during this set of experiments. These parameters were recognized and removed from subsequent model training by the server. Identifying attack models in a timely manner can speed up model convergence and ensure model correctness. Based on the results of the experiments, Fed-SAD demonstrated robustness against adversarial attacks by detecting and removing malicious models during federated learning model aggregation. The study findings highlight the importance of implementing defense strategies against adversarial attacks in the federated learning system.

5 Conclusion

In this paper, we proposed Fed-SAD, a secure aggregation algorithm for federated learning that addresses the problem of malicious models affecting prediction accuracy during load forecast collaborative training. Fed-SAD not only protects data privacy by allowing participants to train their own data collaboratively, but also prevents malicious actors from interfering with the central model training process by securely aggregating models. Experimental simulations demonstrate that our approach is robust against interference from malicious models and can improve prediction accuracy for all participants even in the presence of attacks. This study provides a foundation for further research on secure model aggregation in other distributed load forecast training scenarios, leading to more secure and accurate federated learning processes.

6 Acknowledgments

This work is supported by Research on Key Technologies to Support Network Operation of Distributed Energy Storage (5100-202199544A-0-5-ZN)

7 References

1 L.A Abad, S.M. Sarabia, J.M. Yuzon, and M.C. Pacis. A short-term load forecasting algorithm using support vector regression artificial neural network method (svr-ann). In *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, pages 138–143, 2020.

2 Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

3 Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

4 Coen Bron and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

5 Di Cao, Shan Chang, Zhijian Lin, Guohua Liu, and Donghong Sun. Understanding distributed poisoning attack in federated learning. In *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 233–239. IEEE, 2019.

6 Jiuxiang Chen, Tianlu Gao, Ruiqi Si, Yuxin Dai, Yuqi Jiang, and Jun Zhang. Residential short term load forecasting based on federated learning. In *2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 1–6, 2022.

7 Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *RAID*, pages 301–316, 2020.

8 Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

9 Bai YB Gu YH. Survey on security and privacy of federated learning models. *Journal of Software*, page 2833–2864, 2022.

10 Najeeb Moharram Jebreel, Josep Domingo-Ferrer, Alberto Blanco-Justicia, and David Sánchez. Enhanced security and privacy via fragmented federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

11 Ao Li, Fu Xiao, Cheng Fan, and Maomao Hu. Development of an ann-based building energy model for information-poor buildings using transfer learning. In *Building simulation*, volume 14, pages 89–101. Springer, 2021.

12 Junyang Li, Chaobo Zhang, Yang Zhao, Weikang Qiu, Qi Chen, and Xuejun Zhang. Federated learning-based short-term building energy consumption prediction method for solving the data silos problem. In *Building Simulation*, volume 15, pages 1145–1159. Springer, 2022.

13 Jia Qi Lim and Chee Seng Chan. From gradient leakage to adversarial attacks in federated learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3602–3606. IEEE, 2021.

14 Habib Ullah Manzoor, Muhammed Shahzeb Khan, Ahsan Raza Khan, Fahad Ayaz, David Flynn, Muhammad Ali Imran, and Ahmed Zoha. Fedclamp: An algorithm for identification of anomalous client in federated learning. In *2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 1–4. IEEE, 2022.

15 Clayton Miller and Forrest Meggers. The building data genome project: An open, public data set from non-residential building electrical meters. *Energy Procedia*, 122:439–444, 2017.

16 Arash Moradzadeh, Hamed Moayyed, Behnam Mohammadi-Ivatloo, A. Pedro Aguiar, and Amjad Anvari-Moghaddam. A secure federated deep learning-based approach for heating load demand forecasting in building environment. *IEEE Access*, 10:5037–5050, 2022.

17 Naik Bakht Sania Qureshi, Dong-Hoon Kim, Jiwoo Lee, and Eun-Kyu Lee. On the performance impact of poisoning attacks on load forecasting in federated learning. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, UbiComp '21, page 64–66, New York, NY, USA, 2021. Association for Computing Machinery.

18 Naik Bakht Sania Qureshi, Dong-Hoon Kim, Jiwoo Lee, and Eun-Kyu Lee. Poisoning attacks against federated learning in load forecasting of smart energy. In *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*, pages 1–7. IEEE, 2022.

19 Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

20 Long Su, Zhicong Liu, and Jun Ye. Reputation-based defense scheme against backdoor attacks on federated learning. In *2021 International Conference on Big Data Analytics for Cyber-Physical System in Smart City: Volume 2*, pages 949–955. Springer, 2022.

21 Afaf Taïk and Soumaya Cherkaoui. Electrical load forecasting using edge computing and federated learning. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, 2020.

22 Yuanyuan Wang, Jun Chen, Xiaoqiao Chen, Xiangjun Zeng, Yang Kong, Shanfeng Sun, Yongsheng Guo, and Ying Liu. Short-term load forecasting for industrial customers based on tcn-lightgbm. *IEEE Transactions on Power Systems*, 36(3):1984–1997, 2020.

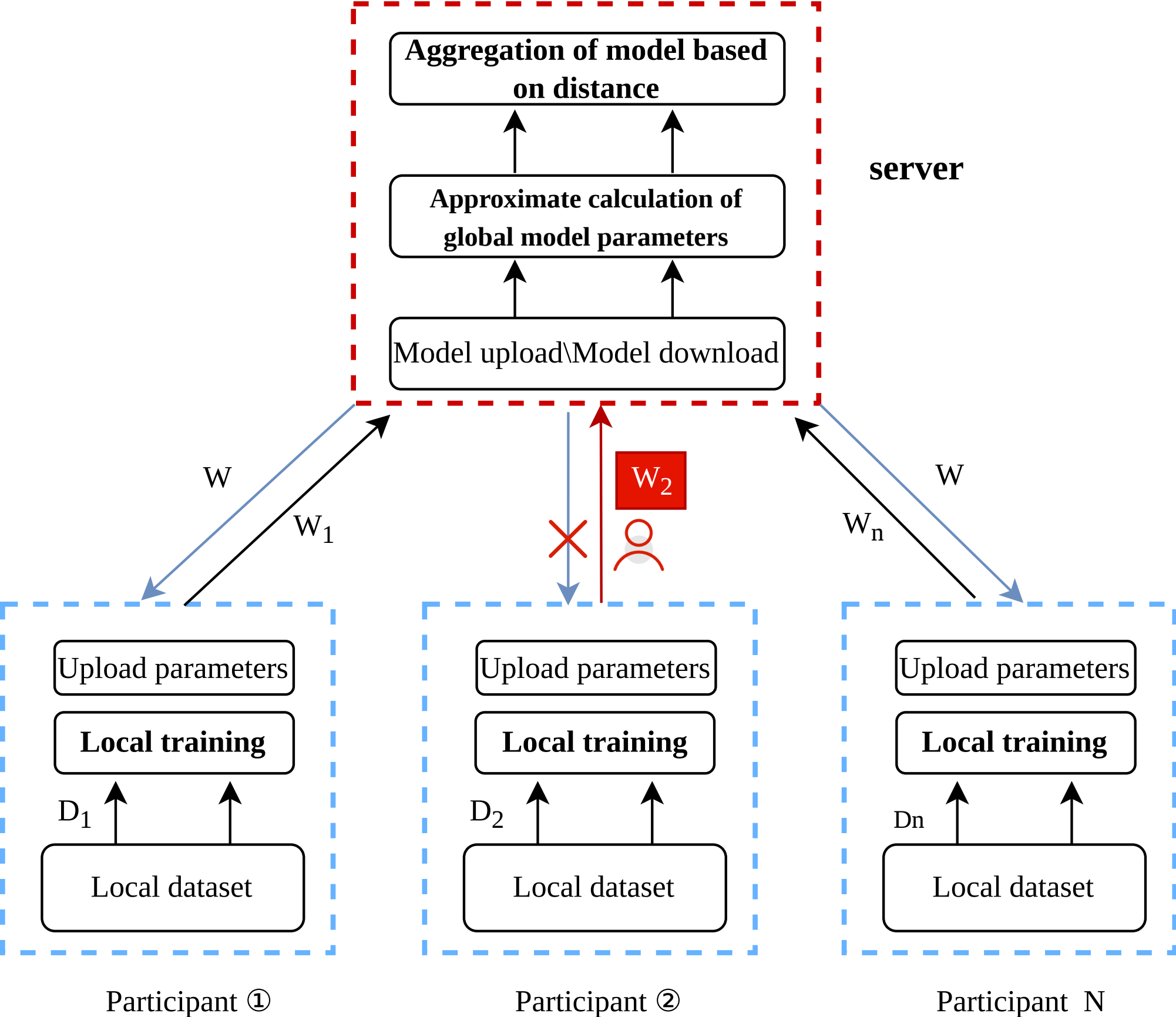
23 Chongchong Xu, Guo Chen, and Chaojie Li. Federated learning for interpretable short-term residential load forecasting in edge computing network. *Neural Comput. Appl.*, 35(11):8561–8574, dec 2022.

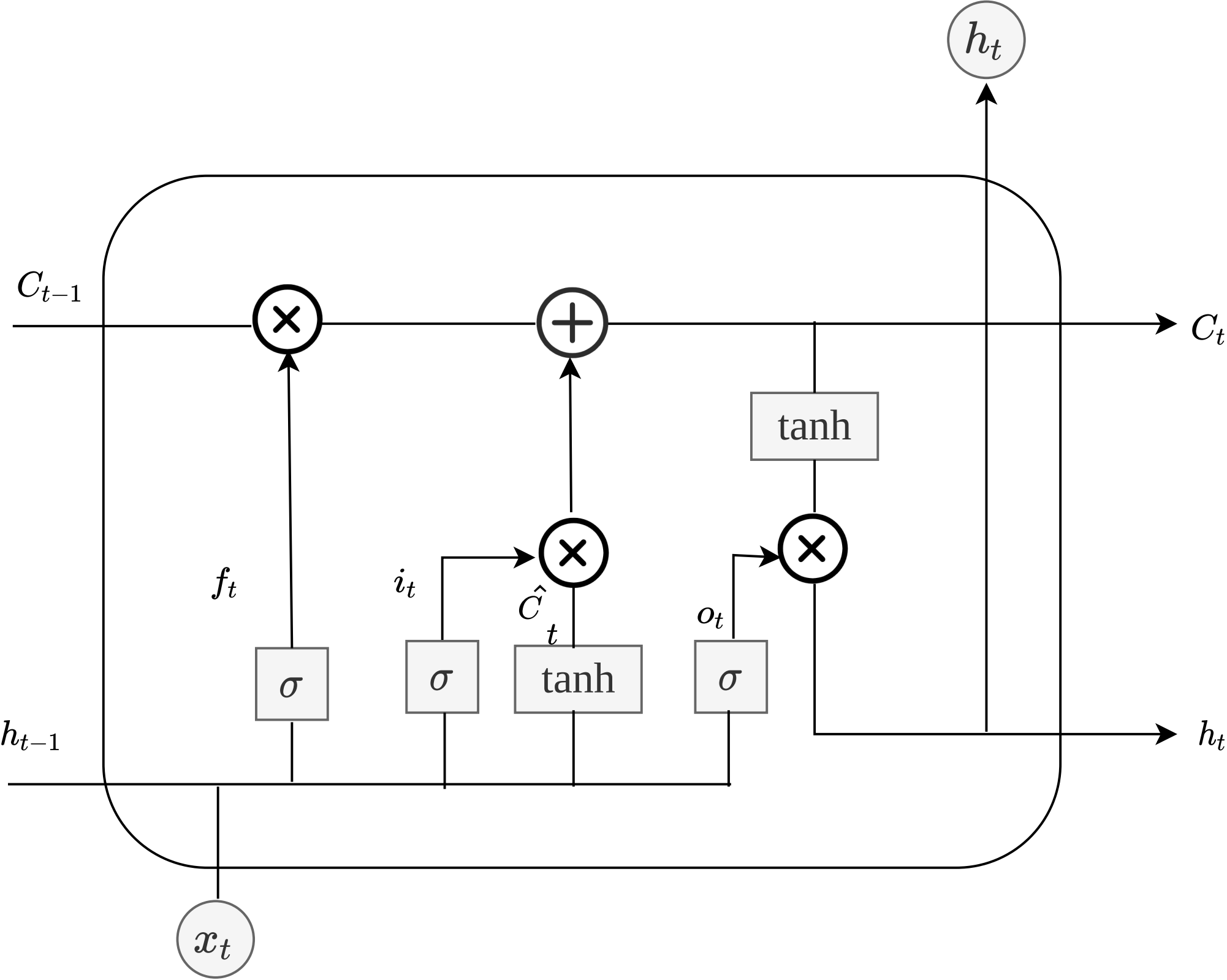
24 Xinyi Xu and Lingjuan Lyu. A reputation mechanism is all you need: Collaborative fairness and adversarial robustness in federated learning. *arXiv preprint arXiv:2011.10464*, 2020.

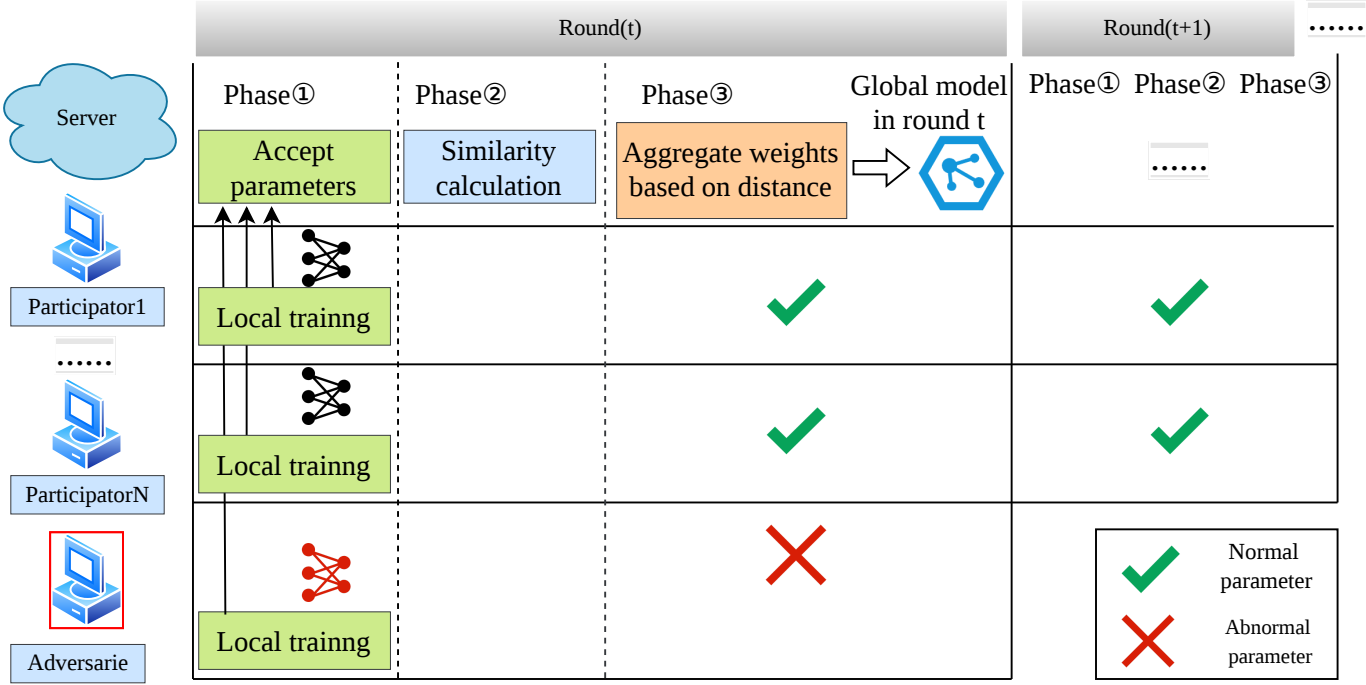
25 Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

26 Qiang Zhang, Quan Yuan, Xin Zhou, and Xin Luo. Research on intelligent load forecast in power system dispatching automation. In *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, pages 575–578, 2021.

27 Chuanxin Zhou, Yi Sun, Degang Wang, and Qi Gao. Fed-fi: Federated learning malicious model detection method based on feature importance. *Security and Communication Networks*, 2022, 2022.


Aggregation of model based on distance
Approximate calculation of global model parameters
Model upload\Model download
server
W
W1
W2
W
Wn
Upload parameters
Local training
D1
Local dataset
Upload parameters
Local training
D2
Local dataset
Upload parameters
Local training
Dn
Local dataset
Participant ①
Participant ②
Participant N


h_t
C_{t-1}
C_t
tanh
f_t
i_t
$\hat{C}_t$
o_t
σ
σ
tanh
σ
h_{t-1}
h_t
x_t


Round(t)
Round(t+1)
Server
Phase①
Phase②
Phase③
Global model in round t
Accept parameters
Similarity calculation
Aggregate weights based on distance
Phase① Phase② Phase③
Participator1
ParticipatorN
Adversarie
Local trainng
Local trainng
Local trainng
Normal parameter
Abnormal parameter