electronics

MDPI

Article

FMDL: Federated Mutual Distillation Learning for Defending Backdoor Attacks

Hanqi Sun[1], Wanquan Zhu [2], Ziyu Sun [3], Mingsheng Cao [4,5] and Wenbin Liu [3,*]

[1] College of Software, Jilin University, Changchun 130012, China; sunhq5521@mails.jlu.edu.cn
[2] School of Information Engineering, Yangzhou University, Yangzhou 225127, China; mx120230578@stu.yzu.edu.cn
[3] College of Computer Science and Technology, Jilin University, Changchun 130012, China; sunzy2121@mails.jlu.edu.cn
[4] Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu 611731, China; cms@uestc.edu.cn
[5] Ningbo WebKing Technology Joint Stock Company, Ltd., Ningbo 315000, China
* Correspondence: liuwenbin@jlu.edu.cn

Abstract: Federated learning is a distributed machine learning algorithm that enables collaborative training among multiple clients without sharing sensitive information. Unlike centralized learning, it emphasizes the distinctive benefits of safeguarding data privacy. However, two challenging issues, namely heterogeneity and backdoor attacks, pose severe challenges to standardizing federated learning algorithms. Data heterogeneity affects model accuracy, target heterogeneity fragments model applicability, and model heterogeneity compromises model individuality. Backdoor attacks inject trigger patterns into data to deceive the model during training, thereby undermining the performance of federated learning. In this work, we propose an advanced federated learning paradigm called Federated Mutual Distillation Learning (FMDL). FMDL allows clients to collaboratively train a global model while independently training their private models, subject to server requirements. Continuous bidirectional knowledge transfer is performed between local models and private models to achieve model personalization. FMDL utilizes the technique of attention distillation, conducting mutual distillation during the local update phase and fine-tuning on clean data subsets to effectively erase the backdoor triggers. Our experiments demonstrate that FMDL benefits clients from different data, tasks, and models, effectively defends against six types of backdoor attacks, and validates the effectiveness and efficiency of our proposed approach.

Keywords: federated learning; heterogeneous; backdoor attack; knowledge distillation; attention map

Citation: Sun, H.; Zhu, W.; Sun, Z.; Cao, M.; Liu, W. FMDL: Federated Mutual Distillation Learning for Defending Backdoor Attacks. *Electronics* **2023**, *12*, 4838. https://doi.org/10.3390/electronics12234838

Academic Editor: Aryya Gangopadhyay

Received: 23 October 2023
Revised: 20 November 2023
Accepted: 21 November 2023
Published: 30 November 2023

1. Introduction

The rapid development of the big data era has highlighted the significant advantages of machine learning in numerous domains, giving rise to a plethora of intelligent applications [1]. However, traditional centralized machine learning suffers from a fatal flaw of highly centralized data, leading to significant privacy breaches. In real-world applications, due to factors such as market competition and management strategies [2], participating users (groups or individuals) are reluctant to share their data due to concerns about privacy risks, thus leading to the problem of data silos. To address this crucial issue, Federated Learning (FL) [3] emerges as a highly promising solution. Its main innovation lies in providing a distributed machine-learning framework with privacy-preserving characteristics, enabling thousands of participants to collaboratively train a specific machine-learning model in a distributed manner. As the training data remains stored locally with the participants throughout the federated learning process, this mechanism allows for the sharing of training data among participants while ensuring privacy protection for each participant [4]. To this day, the improvement and innovation of federated learning frameworks remain a research hotspot in the field of machine learning.

The basic workflow of federated learning is illustrated in Figure 1, which mainly consists of the following steps: (1) Participants download the initialized global model from the cloud server, train the model using their local datasets, and generate the latest local model updates (i.e., model parameters). (2) The cloud server collects the local update parameters and updates the global model using model averaging algorithms. Despite making progress in privacy protection, federated learning still faces numerous security and privacy issues. Among them, the problems of heterogeneity and backdoor attacks are particularly acute.

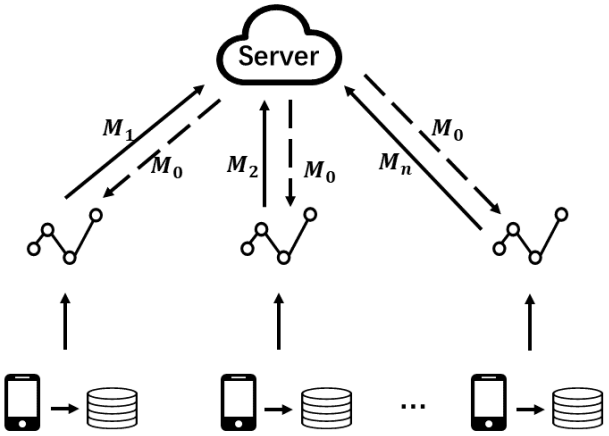


Figure 1. Federated Learning Framework.

Federated learning heterogeneity refers to the differences among various participants in federated learning, which can be summarized into three aspects: data heterogeneity (DH), objective heterogeneity (OH), and model heterogeneity (MH). Data heterogeneity refers to the variations in data characteristics, data types, or data scales among different participants. Each participant may possess data from different domains, such as medical data, financial data, or image data [5], which could exhibit distinct data distributions, feature representations, and data labels. Objective heterogeneity refers to the differences in learning objectives or tasks among different participants. Each participant may have different learning goals, which could include classification, regression, clustering, or other tasks. Model heterogeneity refers to the usage of different machine learning models or model architectures among different participants. Each participant may employ different types of models, such as neural networks, decision trees, support vector machines, and so on. Classical federated learning aims to train a universally applicable global model, which may overfit local data and lose personalized features [6]. Therefore, federated learning requires the design of suitable protocols and algorithms to handle these differences.

Backdoor attacks are not only easy to execute [7] but also possess strong attack capabilities, making them a subject of significant concern in security research for federated learning. A backdoor attack refers to an attempt by an attacker to insert malicious backdoors or traps into a federated learning model. Since federated learning is a distributed machine learning approach involving multiple participants training the model together without sharing raw data, the attacker could be one of the participants or someone attempting to infiltrate the participants. The objective of a backdoor attack is to implant malicious functionality or behavior into the federated learning model, causing the model to perform normally under specific trigger conditions but execute malicious operations under specific backdoor trigger conditions. Examples of triggers include a single pixel [8] or a black-and-white checkerboard [7]. Attackers can implement backdoor attacks by manipulating training data, model parameters, or update rules during the local model update process of the participants. Backdoor attacks pose threats to the security and privacy of federated learning. Once a backdoor is successfully implanted, attackers can exploit the backdoor trigger conditions to perform unauthorized operations or obtain sensitive information. To pre-

vent backdoor attacks, a series of security measures need to be implemented in federated learning, including data privacy protection, participant verification, secure mechanisms for model aggregation, anomaly detection [9], and robustness enhancements.

In this work, we propose a novel federated learning paradigm called Federated Mutual Distillation Learning (FMDL). FMDL is a federated learning strategy that addresses the challenges of federated heterogeneity and backdoor attacks, guided by federated mutual learning and knowledge distillation [10,11]. FMDL views federated learning as a transfer learning process between the global model and local models, using deep mutual learning [12] for localized updates in federated learning. This approach satisfies the universal standards of global updates while preserving the local, personalized requirements. Moreover, under the fine guidance of a teacher model, the student model (i.e., the local model) in FMDL maintains high accuracy even under various backdoor attacks.

We summarize our main contributions as follows:

- We propose the adoption of dual-model updates in local updates of federated learning, where the meme model is used for knowledge transfer between the global model and local models, and the personalized model is designed as a private model for client data and tasks. The two models engage in deep mutual learning to address the three types of heterogeneity, enabling personalized model requirements.
- We construct a clean teacher model based on knowledge distillation to guide the training of the student model. The teacher model is fine-tuned on small, clean subsets to defend against various types of backdoor attacks. This approach significantly reduces the accuracy of backdoor attacks, approaching random guessing without causing significant performance degradation, effectively ensuring privacy and security.
- To achieve defensive performance visualization, we utilize attention maps as an evaluation criterion and define distillation loss based on the attention maps of the teacher and student models.
- We conduct experiments on multiple benchmark datasets to validate the effectiveness of the FMDL method in addressing heterogeneity issues and its security against backdoor attacks.

The remainder of this paper is organized as follows. Section 2 introduces the related works. Section 3 describes the proposed federated mutual distillation learning method. Section 4 evaluates and analyses the results of the experiment. Finally, Section 5 concludes this paper.

2. Related Work

2.1. Federated Learning

Federated Learning coordinates the training of machine learning models across multiple parties while maintaining the privacy of local users. However, it still faces numerous challenges in practice. Highlighting the importance and challenges of group fairness, H. Ezzeldin et al. [13] proposed the FairFed fair-aware aggregation algorithm, which allows the use of debiasing methods across clients and demonstrates advantages in scenarios with highly heterogeneous client data. Zhang et al. [14] introduced FedALA (Adaptive Local Aggregation) to improve the generalization ability of the global model by capturing the information required from client models in personalized federated learning. Similarly, Huang et al. [15] proposed Federated Prototype Learning (FPL), which constructs cluster prototypes and unbiased prototypes to provide rich domain knowledge and fair convergence objectives. Simple federated learning often requires a large number of training iterations to converge and lacks adaptability. To address this issue, Wu et al. [16] designed an efficient adaptive algorithm called FAFED (Fast Adaptive Federated Learning) based on momentum-based variance reduction techniques in cross-silo federated learning, significantly improving the efficiency of heterogeneous data in language modeling tasks and image classification tasks.

2.2. Heterogeneous Federated Learning

The heterogeneity of data in federated learning can be attributed to the generation of data from various clients with different distributions. Federated learning achieves high model accuracy when trained on identically distributed data. However, non-IID data can lead to imbalanced data distributions, introducing bias during model training due to weight differences during model aggregation. As a result, the performance of federated learning may significantly decline. To address this, Zhao et al. [17] proposed a data-sharing strategy that creates shared data subsets to mitigate non-IID data. Wang et al. [18] utilized distillation to extract shared data, while Chen et al. [19] employed generative adversarial networks to generate shared data, both achieving promising results. However, implementing shared data methods can be complex.

Objective heterogeneity in federated learning refers to the differences between the global model and local model objectives. The central server performs global aggregation to ensure participants obtain a generalized global model and converge iteratively toward a universally applicable model. On the other hand, clients train local models using private data to obtain representative personalized private models. Clients also expect their models to perform well in the global model after each local model update. However, due to data heterogeneity, federated aggregation [3] sacrifices some clients' individualities in favor of commonalities, thereby excluding these clients from benefiting from federated learning [20]. Liu et al. [21] proposed a federated learning framework where different tasks yield multiple types of image representations, and useful features from vision and language are merged to enhance personalized models.

Due to the variations in client hardware capabilities [22], different representations of local data [12,23], and diverse client task requirements [24], clients often need to individually design their private models, resulting in model heterogeneity in federated learning. Khodak et al. [25] proposed the ARUBA theoretical framework, which allows individually trained models controlled by a central server but lacks a comprehensive implementation method. Li et al. [26] proposed a decentralized framework that utilizes knowledge distillation to enable federated learning with independently designed models, but this method does not support new participants.

2.3. Knowledge Distillation

Knowledge distillation is a model compression technique aimed at transferring knowledge from a large, complex model, referred to as the teacher model, to a smaller, simplified model known as the student model. The goal is to reduce model complexity, improve inference speed, and decrease model storage space. The teacher model is characterized by its large scale, strong performance, and abundant knowledge, exhibiting high accuracy and expressive power, and it guides the student model, which is smaller, simplified, and has weaker knowledge. Knowledge distillation has shown great potential in various aspects such as adversarial robustness [27], multi-granular lip-reading [28], and data augmentation [29]. To supervise the training of the student model and improve distillation performance for better model performance [30,31], feature maps and attention maps [32–34] are widely utilized as the basis for visual analysis.

2.4. Backdoor Attacks and Defense

Backdoor attacks aim to inject triggers (poison labels [35,36] or clean labels [37–39]) into a small portion of the model data during training to disrupt model predictions and compromise model performance, achieving the desired attack effect. Li et al. [40] investigated various aspects of backdoor attacks. In addition to the single-pixel and black-box modules mentioned earlier, real-world backdoor attacks are more covert, and there are six common and more advanced attack methods: BadNets [7], Trojan attack [36], Blend attack [35], Clean-label attack (CL) [38], Sinusoidal signal attack (SIG) [41], and Reflection attack (Refool) [39].

When dealing with these attacks, the backdoor defense can be summarized into two approaches: backdoor detection and trigger erasing. Detection-based methods can only verify the presence of a backdoor and whether the model is poisoned while the backdoor still remains in the model, rendering some data or the entire model unusable. Erasing-based methods can effectively remove the backdoor and purify the model. Liu et al. [42] utilized fine-grained pruning to remove the implanted backdoor information neurons and fine-tuned the model to suppress backdoor triggers. However, fine-tuning methods can lead to a decrease in model performance as the model may become overly adapted to clean subsets [43]. Truong et al. [44] proposed regularization-based methods, and Zhao et al. [45] introduced pattern stitching for backdoor repair, but the effectiveness of these methods often does not outweigh the costs, as effective methods often come at a high price. In the latest research on backdoors, Jebreel et al. [46] proposed a defense mechanism called FL-Defengder, which filters poisoned samples by analyzing the feature differences between key-layer suspicious samples and benign samples. They further addressed FL-targeted attacks by updating the centroid bias of similar vectors obtained through re-weighting client-side PCA compression. This approach achieves a lower attack success rate while maintaining task accuracy [47].

3. Federated Mutual Distillation Learning

Addressing the three heterogeneous issues in federated learning while maintaining model personalization and improving overall performance, we introduce the method of knowledge distillation during the local update phase. Unlike the typical teacher-student relationship in knowledge distillation structures, although there is no well-trained teacher model or untrained student model in the federated learning system, the method of knowledge distillation can be applied to two different knowledge transfer models with different architectures. Therefore, we deviate from the traditional concept of one-way knowledge transfer in the "teacher-student model" and employ deep mutual learning in federated learning for local model updates. This allows the central server to obtain a generalized global model while enabling different clients to train a private personalized model tailored to their specific data and task requirements, resulting in a win-win situation. As shown in Figure 2, we design two types of model structures within each client: (1) a local model used to receive the global model for local training and updates, and (2) a private model designed by the client for their specific needs. Both models engage in continuous mutual learning.

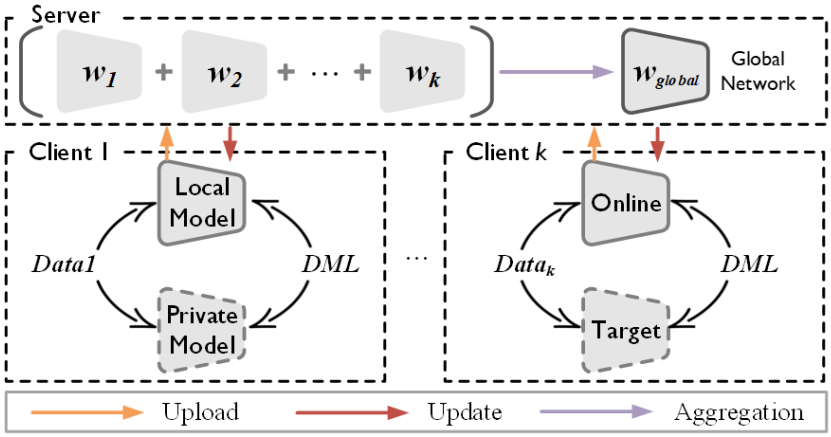


Figure 2. Federated Mutual Distillation Learning.

Additionally, we report our experimental parameters in Table 1.

Table 1. Symbolic representations used in FMDL.

Symbols	Meaning
L	Loss Function.
L_{CE}	Cross Entropy.
$D_{KL}(\|\|)$	Kullback Leibler (KL) Divergence.
p	The predicts of the model.
T	A hyper-parameter mean temperature.
z	The logits of teacher model.
α, β	The hyper-parameters that control the proportion of knowledge from data or from the other model.
M	A DNN model.
M^l	The activation output at the l-th layer.
Z	An attention operator that maps an activation map to an attention representation.
Z_{sum}	Reflects all activation regions.
Z^k_{sum}	Amplify the disparities between the backdoored neurons and the benign neurons by an order of p.
Z^k_{mean}	Align the activation center of the backdoored neurons with that of the benign neurons by taking the mean over all activation regions.
N, H, W	The dimensions of the channel, the height, and the width of the activation map.

3.1. Classical Distillation Methods

The fundamental idea behind knowledge distillation is to transfer the "knowledge" of the teacher model to the student model by having the student model learn the teacher model's predicted outcomes. Through knowledge distillation, the student model can acquire additional information from the teacher model, including relationships between categories, decision boundaries, and data distributions. As a result, the student model can maintain relatively high accuracy while having a smaller model size and faster inference speed. The loss function for the student model can be simplified as follows:

$$L_{student} = L_{CE} + D_{KL}(p_{teacher}||p_{student}) \tag{1}$$

$$p_{teacher} = \frac{exp(z/T)}{\sum_i exp(z_i/T)} \tag{2}$$

3.2. Federated Mutual Learning

In traditional federated learning, each participant (or client) trains a local model using their own local data and only shares model updates with the central server. The central server aggregates these updates to create a global model, which is then distributed back to the participants. However, the collaboration among participants is limited to the exchange of model updates. In terms of target heterogeneity, typical federated learning only focuses on the objectives of the central server and overlooks the clients' need for personalized models. Moreover, in cases of significant data heterogeneity, the performance of the globally trained model in federated learning may not be ideal. In fact, if the client's data are fragmented and dispersed, the model may even fail to converge after multiple iterations. Therefore, we introduce federated mutual learning, which aims to leverage the collective intelligence and diverse perspectives of participants to enhance the learning process, improve model performance, and overcome the individual limitations of participant data.

During the training process of federated mutual learning, the initial model is still distributed by the central server and used as the local model for the first iteration of local updates. Simultaneously, all clients also customize an independent private model (allowing for similarity or diversity). Both models are trained using local data. Unlike normal local updates, each client does not simply train a replica of the global model but instead engages in several rounds of deep mutual learning between the local model and the private model to achieve better performance than independent training. The detailed process is illustrated in

Figure 3. During this process, knowledge is transferred bidirectionally. As the local model receives updates from the global model, it migrates the knowledge from the central server to the private model. At the same time, the private model provides feedback on the client's personalized features. Finally, the client sends the trained local model to the server, which selects and aggregates them for updating the global model in the federated aggregation step, preparing for the next round of federated training. This process is repeated until the model converges. The objectives of both models are to undergo self-training on the same dataset to achieve consistency in the predicted outcomes. The entire process is illustrated in Algorithm 1. Throughout the training process, we redefine the classical knowledge distillation loss as follows:

$$L_{private} = \alpha L_{C_{private}} + (1-\alpha) D_{KL}(p_{local} \parallel p_{private}) \tag{3}$$

$$L_{local} = \beta L_{C_{local}} + (1-\beta) D_{KL}(p_{private} \parallel p_{local}) \tag{4}$$

Algorithm 1: Federated Mutual Learning

Global Updata:
Distribute the initial global model G_0.;
for *epoch* $t = 1, 2, \cdots, N$ **do**
  for *client* $i = 1, 2, \cdots, N$ **do**
    $Local_{t+1}^{i} \leftarrow LocalUpdata(Local_t^i)$;
  Aggregation:$G_{t+1} \leftarrow \frac{1}{N} \sum_{i=1}^{N} Local_{t+1}^{i}$;
LocalUpdata:
Initialize the private personalized model $private_0^i$.;
$Local_t^i \leftarrow G_t$;
for *epoch* $t = 1, 2, \cdots, T$ **do**
  DML between local and private models over private data.;

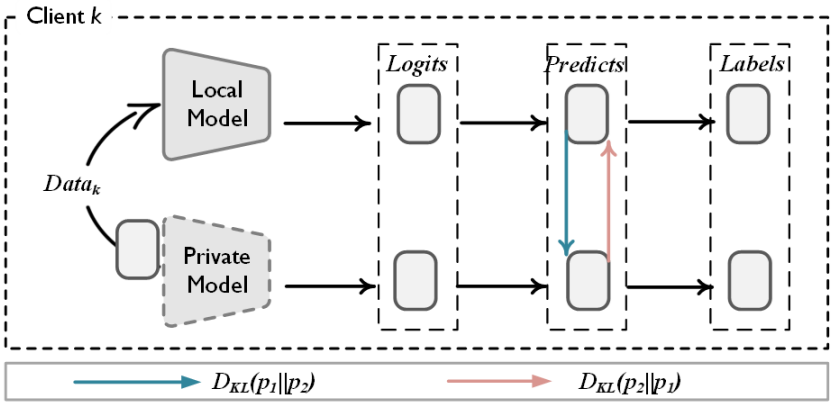


Figure 3. Federated Mutual Learning.

3.3. Attention Distillation Defense Methods

Attention maps are commonly used in deep learning to visualize the regions or positions that the model focuses on when processing inputs. Attention mechanisms are employed in sequence data tasks such as natural language processing and computer vision tasks. The attention mechanism calculates weights for each input element, allowing the model to concentrate on the most relevant or important elements. The attention map provides a visual representation of these weight allocations. Given a deep neural network model M, we define $Z\colon \mathbb{R}^{N\times H\times W} \rightarrow \mathbb{R}^{H\times W}$ as the attention algorithm that maps the activation maps to the attention representation, i.e., transforming the 3D activation maps into a flattened 2D tensor along the channel dimension. Attention maps play a crucial role

in successful knowledge distillation. There are three common forms of attention algorithms, as shown below:

$$Z_{sum}(M^l) = \sum_{i=1}^{N} |M_i^l|; Z_{sum}^p(M^l) = \sum_{i=1}^{N} |M_i^l|^p; Z_{mean}^p(M^l) = \frac{1}{N} \sum_{i=1}^{N} |M_i^l|^p \tag{5}$$

Since the teacher model and student model in knowledge distillation do not have a direct correspondence in federated learning, we replace the "teacher-student relationship" with the view that the local model and private model can be considered as two different student models. They engage in mutual learning and distillation, forming a "student-student relationship". As the local model continuously participates in federated learning iterations, there may be potential backdoors. The process of erasing backdoor triggers through attention distillation is illustrated in Figure 4. Therefore, it is necessary to add attention loss. Attention distillation combines the local model and private model through a neural attention extraction process. Attention representations are computed after each residual block, and attention distillation loss is defined based on the attention representations of both models:

$$L_{AD}(M_{local}^l, M_{private}^l) = \| \frac{Z(M_{local}^l)}{\| Z(M_{local}^l) \|_2} - \frac{Z(M_{private}^l)}{\| Z(M_{private}^l) \|_2} \|_2 \tag{6}$$

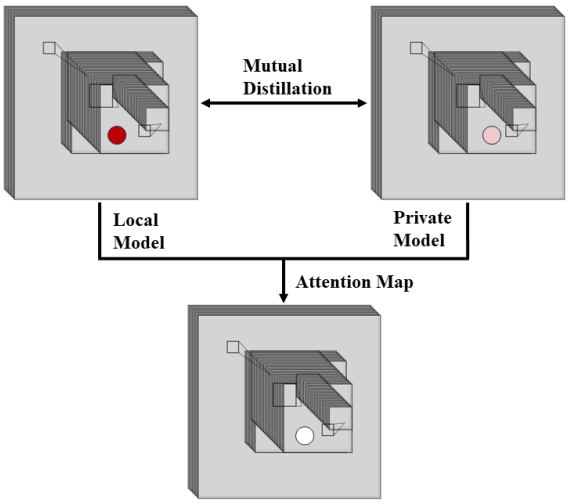


Figure 4. Attention distillation.The red dot in the figure represent the embedded backdoors in the model, the pink dot represent the backdoors after fine-tuning, and the white dot represent the removed backdoors.

Therefore, the overall training loss can be expanded as:

$$L_{total} = \beta L_{C_{local}} + (1-\beta) D_{KL}(p_{private} \parallel p_{local}) + \sum_{l=1}^{P} L_{AD}(M_{local}^l, M_{private}^l) \tag{7}$$

4. Experimental Evaluations

In this section, we conducted comparative experiments between the proposed Federated Mutual Learning (FMDL) method and the traditional Federated Learning (FL) method. The performance of FMDL was evaluated on three commonly used image classification datasets. Additionally, we conducted experiments specifically targeting three types of heterogeneous problems. In terms of backdoor defense, we assessed the performance of FMDL compared to three existing defense methods based on erasure techniques under six common backdoor attacks. Moreover, we clarified the criteria for selecting the form of attention maps representation.

4.1. Experimental Setup

We utilized three datasets, namely MNIST, CIFAR-10, and CIFAR-100, for training in federated learning, as shown in Table 2. MNIST dataset: this dataset contains 70,000 handwritten digital images of 10 classes (0–9), including 60,000 training samples and 10,000 test samples. In all experiments, samples from the MNIST dataset were normalized to 28 × 28 pixels. For the pixel block backdoor attack, the attacker embeds a 5 × 5 pixel block in the samples and assigns them with the target label "1" desired by attackers. In the watermarking backdoor attack experiment, the attacker added the watermarking "1" with a different watermarking factor to some real samples and set its label as "1". CIFAR-10 dataset: the CIFAR-10 dataset contains 60,000 color images in 10 categories (such as "aircraft", "car", "bird", etc.), including 50,000 training samples and 10,000 test samples. The images in the CIFAR-10 dataset are normalized to a 32 × 32 three-channel input during data preprocessing. For the attribute backdoor attack experiment, "car" in the CIFAR-10 dataset, "cars with stripes", "cars next to striped walls", and "green cars" were selected as the attribute backdoor, which the backdoor triggers. CIFAR-100 has the same total number of images as CIFAR-10, but it has 100 classes. CIFAR-100 has 500 training images and 100 testing images per class. Furthermore, a Multilayer Perceptron (MLP) model was employed, where the weights and biases between neurons were updated using the backpropagation algorithm to minimize the loss function. After training, the data were classified. A Convolutional Neural Network (CNN) was used to extract different features by generating convolutional feature maps with 3 × 3 convolutional kernels. ReLU activation was applied to two convolutional layers (the first layer with 6 channels and the second layer with 16 channels, both followed by 2 × 2 max pooling). The linear layer and softmax layer were utilized for output. The optimizer chosen was the Stochastic Gradient Descent (SGD) algorithm with momentum = 0.9, weight decay = 5×10^{-4}, and batch size = 128.

Table 2. Datasets used in our experiments.

Dataset	Training Samples	Test Samples	Classes	Model
MNIST	60,000	10,000	10	MLP
CIFAR-10	50,000	10,000	10	CNN
CIFAR-100	50,000	10,000	100	CNN

The selection of the distillation parameter β is crucial for clearing the backdoor. Intuitively, a larger β is more effective in defending against backdoors. However, arbitrarily increasing the value of β may lead to a decline in the performance of the method. Based on the scaling experiments in [48], we have determined the value of β to be 0.5. Although increasing β always enhances model robustness, setting β to 0.5 has already reduced the clean accuracy below the threshold, which is the optimal value of β.

To ensure fair evaluation, we followed the experimental configurations of six backdoor attacks as described in their respective original papers, including trigger models, sizes, and target labels, as presented in Table 3. Regarding the backdoor defense methods, we compared three methods: Fine-tuning, Fine Pruning, and Mode Connectivity Repair (MCR), with our proposed FMDL method. For all defense methods, we assumed access to 5% of clean data.

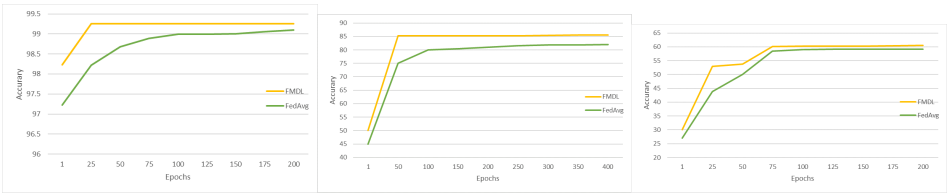
We used two metrics to evaluate the performance of the defense mechanisms: Attack Success Rate (ASR) and Accuracy on Clean Samples (ACC). A higher decrease in ASR and a lower decrease in ACC indicate a stronger defense mechanism.

Table 3. Experimental configurations of six backdoor attacks.

Backdoor	BadNets	Trojan	Blend	Clean-Label	Signal	Refool
Dataset	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-10	GTSRB
Model	WideResNet	WideResNet	WideResNet	WideResNet	WideResNet	WideResNet
Inject Rate	0.1	0.05	0.1	0.08	0.08	0.08
Trigger Type	Grid	Square	Random Noise	Grid+PGD Noise	Sinusoidal Signal	Reflection
Target Size	3×3	3×3	Full Image	3×3	Full Image	Full Image

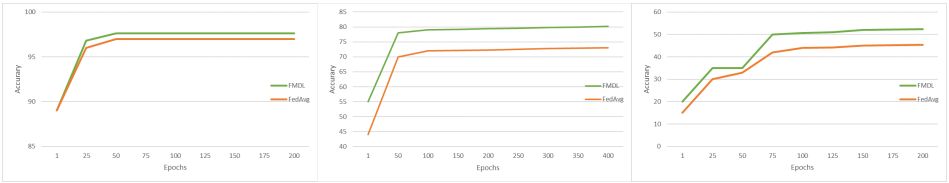
4.2. Comparison of FMDL and Traditional FL Performance

To test the basic performance of FMDL, i.e., whether it can train a universal and effective global model similar to classical federated learning, we conducted comparative experiments between FMDL and FL after the FedAvg procedure. We evaluated the performance of both methods on three different datasets. Additionally, we constructed different data structures for each dataset, namely IID data (as shown in Figure 5) and Non-IID data (as shown in Figure 6). Based on the performance of FMDL on different datasets, we can conclude that our proposed method outperforms traditional federated learning in various aspects. Compared to FedAvg, FMDL demonstrates advantages in terms of faster convergence, higher accuracy, and model stability across different dataset structures.



(**a**) (**b**) (**c**)

Figure 5. The correlation between accuracy and training epochs on IID data. (**a**) MNIST; (**b**) CIFAR-10; (**c**) CIFAR-100.



(**a**) (**b**) (**c**)

Figure 6. The correlation between accuracy and training epochs on non-IID data. (**a**) MNIST; (**b**) CIFAR-10; (**c**) CIFAR-100.

4.3. Performance of FMDL under Three Heterogeneous Settings

When comparing the corresponding (a), (b), and (c) subfigures in Figures 5 and 6, it can be observed that traditional federated learning achieves significantly lower model accuracy when trained on Non-IID data compared to IID data for the CIFAR-10 or CIFAR-100 datasets, as depicted in the figures. This is due to data heterogeneity causing imbalanced weights during local updates, and simple federated aggregation hindering model progress. In contrast, FMDL maintains higher accuracy for both IID and non-IID data. Although there is a slight decrease in accuracy for non-IID data compared to IID data, it remains within an acceptable range. Data heterogeneity has a consistent impact on global performance, which aligns with real-world scenarios. On the MNIST dataset, the model achieves near-perfect predictions, indicating the effectiveness of FMDL in addressing data heterogeneity.

FMDL successfully achieves the central server's objective of training a well-performing generalized model for target heterogeneity. Regarding personalized requirements, our designed mutual learning structure ensures that private models are trained locally without participating in global and local model updates. As a result, the private models fully satisfy the client's needs, achieving model personalization.

We conducted experiments with FMDL using five participating clients. Initially, we independently trained the five clients to obtain personalized models with the best accuracy, as indicated by the yellow portion in Figure 7. Subsequently, we trained all clients using FMDL, and the accuracy of the personalized models obtained is shown in the orange portion of Figure 7. Comparative analysis reveals that the accuracy of the models obtained using our proposed method is higher than that of individually trained models. This demonstrates that FMDL enables clients with different models to benefit from a shared model, effectively addressing model heterogeneity.

Figure 7. Performance of FMDL under the Model Heterogeneous.

4.4. Effectiveness of FMDL in Defending Backdoors

In the previous section, we proposed three representations for attention functions and conducted the following experiments to identify the functions that exhibit better performance. Using the BadNet attack as the baseline attack and ASR (Attack Success Rate) and ACC (Accuracy) as evaluation metrics, we obtained the following results in Table 4. Hence, we adopted Z^2_{sum} as our computational function for calculating the overall distillation loss of the model.

Table 4. The best results of different attention functions.

Attention Function	Z_{mean}		Z^2_{mean}		Z_{sum}		Z^2_{sum}	
	ACC	**ASR**	**ACC**	**ASR**	**ACC**	**ASR**	**ACC**	**ASR**
Baseline	100%	85.86%	100%	85.86%	100%	85.86%	100%	85.86%
Epoch 5	12.28%	81.50%	4.60%	81.30%	6.89%	81.46%	4.21%	81.55%

Firstly, we attacked the model using six different backdoor attacks and employed four different defense mechanisms to evaluate their respective attack success rates. Next, we tested the accuracy of the backdoor models after erasing them on clean samples. The results are shown in Table 5. The MCR (Masking and Confusion Rule) defense method performed remarkably well in countering BadNets and SIG attacks, resulting in the lowest backdoor attack success rate. However, it showed mediocre performance against other attacks. On the other hand, the Fine-tuning method exhibited relatively good prediction results on clean samples after BadNets and CL attacks, but the success rate of backdoor attacks did not decrease significantly, making it an inadequate defense mechanism. In comparison to the other three methods, our proposed attention distillation defense method demonstrated excellent performance in reducing the accuracy of multiple backdoor attacks. Simultaneously, the model's prediction accuracy on clean samples did not suffer significant losses, with an

average deviation of 2.66%, which is within an acceptable range. The attention distillation method showcased effectiveness and efficiency in countering backdoor attacks.

Table 5. Performance of our method against six backdoor attacks.

Backdoor Attack	BadNets	Trojan	Blend	CL	SIG	Refool
Before ASR	100	100	99.97	99.21	99.91	95.16
ACC	85.65	81.24	84.95	84.95	84.36	82.38
AD ASR	4.77	19.63	4.04	9.18	2.52	3.18
ACC	81.17	79.16	81.68	80.34	81.95	80.73

We also verified the impact of different proportions of clean samples on model performance, as shown in Figure 8. According to the information reflected in the bar chart, when the proportion of clean samples reached 20%, both the MCR method and our proposed FMDL method reduced the backdoor attack success rate to below 5%. However, our proposed method exhibited better convergence speed than the MCR method. Even with only 1% of clean samples, FMDL reduced the average ASR from 99.04% to 35.93%, while MCR had a high attack success rate of 80%.

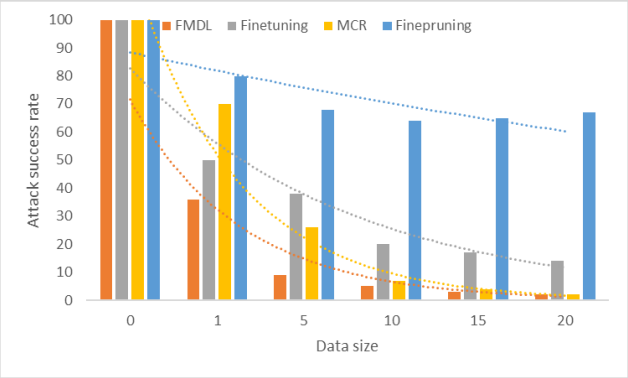


Figure 8. The impact of different proportions of clean samples.

5. Conclusions

Considering the challenges of federated learning in three heterogeneous settings and backdoor attacks, this paper proposes a knowledge distillation-based federated learning paradigm called Federated Mutual Distillation Learning (FMDL). In the local update phase, we introduce mutual learning and mutual distillation between local models and private models to address heterogeneity, and experimental results demonstrate its effectiveness. Additionally, FMDL employs attention maps to evaluate the performance of defense mechanisms. The results show that our proposed method outperforms three other backdoor defense methods in countering six backdoor attacks. Overall, our FMDL method makes significant contributions to addressing heterogeneous federated learning and mitigating the threat of backdoor attacks in model deployment. Future research will explore more advanced methods to achieve federated personalization. Furthermore, although the distillation-based approach effectively eliminates backdoors, the reliance on teacher models increases the computational burden on clients. In practical scenarios, users may opt for less computationally expensive methods with slightly lower performance to mitigate backdoors. Therefore, exploring methods to reduce computational overhead is worth investigating. Additionally, the attention map we utilized lacks strict theoretical analysis, and there is a lack of mature theoretical analysis tools for backdoor attacks. Hence, it is crucial to explore theoretical analysis methods for backdoor attacks.

Author Contributions: Conceptualization, H.S. and W.Z.; methodology, H.S.; software, Z.S.; validation, W.Z., Z.S. and M.C.; formal analysis, Z.S.; data curation, W.Z.; writing—original draft preparation, H.S.; writing—review and editing, W.L.; visualization, Z.S.; supervision, M.C.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No. 62102161, 62002047).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: Author Mingsheng Cao was employed by the company Ningbo WebKing Technology Joint Stock Company, Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ribeiro, M.; Grolinger, K.; Capretz, M.A. Mlaas: Machine learning as a service. In Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 9–11 December 2015; pp. 896–902.
2. Ren, Y.; Leng, Y.; Qi, J.; Sharma, P.K.; Wang, J.; Almakhadmeh, Z.; Tolba, A. Multiple cloud storage mechanism based on blockchain in smart homes. *Future Gener. Comput. Syst.* **2021**, *115*, 304–313. [CrossRef]
3. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
4. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [CrossRef]
5. Ren, Y.; Leng, Y.; Cheng, Y.; Wang, J. Secure data storage based on blockchain and coding in edge computing. *Math. Biosci. Eng* **2019**, *16*, 1874–1892. [CrossRef] [PubMed]
6. Jiang, Y.; Konečnỳ, J.; Rush, K.; Kannan, S. Improving federated learning personalization via model agnostic meta learning. *arXiv* **2019**, arXiv:1909.12488.
7. Gu, T.; Dolan-Gavitt, B.; Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv* **2017**, arXiv:1708.06733.
8. Tran, B.; Li, J.; Madry, A. Spectral signatures in backdoor attacks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
9. Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; Srivastava, B. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv* **2018**, arXiv:1811.03728.
10. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541.
11. Urban, G.; Geras, K.J.; Kahou, S.E.; Aslan, O.; Wang, S.; Caruana, R.; Mohamed, A.; Philipose, M.; Richardson, M. Do deep convolutional nets really need to be deep and convolutional? *arXiv* **2016**, arXiv:1603.05691.
12. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
13. Ezzeldin, Y.H.; Yan, S.; He, C.; Ferrara, E.; Avestimehr, A.S. Fairfed: Enabling group fairness in federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 7494–7502.
14. Zhang, J.; Hua, Y.; Wang, H.; Song, T.; Xue, Z.; Ma, R.; Guan, H. FedALA: Adaptive local aggregation for personalized federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 11237–11244.
15. Huang, W.; Ye, M.; Shi, Z.; Li, H.; Du, B. Rethinking federated learning with domain shift: A prototype view. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 16312–16322.
16. Wu, X.; Huang, F.; Hu, Z.; Huang, H. Faster adaptive federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 10379–10387.
17. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated learning with non-iid data. *arXiv* **2018**, arXiv:1806.00582.
18. Wang, T.; Zhu, J.Y.; Torralba, A.; Efros, A.A. Dataset distillation. *arXiv* **2018**, arXiv:1811.10959.
19. Chen, H.; Wang, Y.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Xu, C.; Xu, C.; Tian, Q. Data-free learning of student networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3514–3522.
20. Yu, T.; Bagdasaryan, E.; Shmatikov, V. Salvaging federated learning by local adaptation. *arXiv* **2020**, arXiv:2002.04758.
21. Liu, F.; Wu, X.; Ge, S.; Fan, W.; Zou, Y. Federated learning for vision-and-language grounding problems. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11572–11579.

22. Wu, B.; Dai, X.; Zhang, P.; Wang, Y.; Sun, F.; Wu, Y.; Tian, Y.; Vajda, P.; Jia, Y.; Keutzer, K. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10734–10742.
23. Gao, D.; Ju, C.; Wei, X.; Liu, Y.; Chen, T.; Yang, Q. Hhhfl: Hierarchical heterogeneous horizontal federated learning for electroencephalography. *arXiv* **2019**, arXiv:1909.05784.
24. Smith, V.; Chiang, C.K.; Sanjabi, M.; Talwalkar, A.S. Federated multi-task learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4427–4437.
25. Khodak, M.; Balcan, M.F.F.; Talwalkar, A.S. Adaptive gradient-based meta-learning methods. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5917–5928.
26. Li, D.; Wang, J. Fedmd: Heterogenous federated learning via model distillation. *arXiv* **2019**, arXiv:1910.03581.
27. Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2016; pp. 582–597.
28. Zhao, Y.; Xu, R.; Wang, X.; Hou, P.; Tang, H.; Song, M. Hearing lips: Improving lip reading by distilling speech recognizers. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 6917–6924.
29. Bagherinezhad, H.; Horton, M.; Rastegari, M.; Farhadi, A. Label refinery: Improving imagenet classification through label progression. *arXiv* **2018**, arXiv:1805.02641.
30. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
31. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
32. Song, X.; Feng, F.; Han, X.; Yang, X.; Liu, W.; Nie, L. Neural compatibility modeling with attentive knowledge distillation. In Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 5–14.
33. Ahn, S.; Hu, S.X.; Damianou, A.; Lawrence, N.D.; Dai, Z. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9163–9171.
34. Heo, B.; Lee, M.; Yun, S.; Choi, J.Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; Volume 33, pp. 3779–3787.
35. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv* **2017**, arXiv:1712.05526.
36. Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.C.; Zhai, J.; Wang, W.; Zhang, X. Trojaning attack on neural networks. In Proceedings of the 25th Annual Network And Distributed System Security Symposium (NDSS 2018), San Diego, CA, USA, 18–21 February 2018.
37. Shafahi, A.; Huang, W.R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; Goldstein, T. Poison frogs! Targeted clean-label poisoning attacks on neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
38. Turner, A.; Tsipras, D.; Madry, A. Clean-Label Backdoor Attacks. 2018. Available online: https://openreview.net/forum?id=HJg6e2CcK7 (accessed on 20 November 2023).
39. Liu, Y.; Ma, X.; Bailey, J.; Lu, F. Reflection backdoor: A natural backdoor attack on deep neural networks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part X 16; pp. 182–199.
40. Li, Y.; Jiang, Y.; Li, Z.; Xia, S.T. Backdoor learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
41. Barni, M.; Kallas, K.; Tondi, B. A new backdoor attack in cnns by training set corruption without label poisoning. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 101–105.
42. Liu, K.; Dolan-Gavitt, B.; Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*; Springer: Cham, Switzerland, 2018; pp. 273–294.
43. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef] [PubMed]
44. Truong, L.; Jones, C.; Hutchinson, B.; August, A.; Praggastis, B.; Jasper, R.; Nichols, N.; Tuor, A. Systematic evaluation of backdoor data poisoning attacks on image classifiers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 788–789.
45. Zhao, P.; Chen, P.Y.; Das, P.; Ramamurthy, K.N.; Lin, X. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv* **2020**, arXiv:2005.00060.
46. Jebreel, N.M.; Domingo-Ferrer, J. FL-Defender: Combating targeted attacks in federated learning. *Knowl.-Based Syst.* **2023**, *260*, 110178. [CrossRef]
47. Jebreel, N.M.; Domingo-Ferrer, J.; Li, Y. Defending Against Backdoor Attacks by Layer-wise Feature Analysis. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Osaka, Japan, 25–28 May 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 428–440.
48. Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; Ma, X. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv* **2021**, arXiv:2101.05930.