EarlyShield: Early-Stage Screening for Robust Personalized Federated Learning

Shixiong Li[1], Xingyu Lyu[1], Ning Wang[2], Tao Li[3], Danjue Chen[4], Yidan Hu[5], and Yimin Chen[1]

[1] Miner School of Computer and Information Sciences, UMass Lowell, USA
{shixiong_li, xingyu_lyu, ✉ian_chen}@uml.edu
[2] Department of Computer Science and Engineering, University of South Florida, USA
ningw@usf.edu
[3] Department of Computer and Information Technology, Purdue University, USA
li4270@purdue.edu
[4] Department of Civil, Construction, and Environmental Engineering, NC State University, USA
dchen33@ncsu.edu
[5] Department of Cybersecurity, Rochester Institute of Technology, USA
yidan.hu@rit.edu

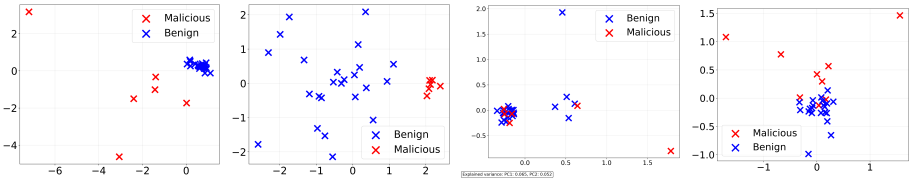
Abstract. Backdoor attacks pose a serious threat to federated learning (FL). The challenge becomes even more pronounced in personalized FL (PFL), where model updates naturally exhibit high diversity across clients. Existing defenses such as clustering-based detection fail under PFL because benign updates appear highly heterogeneous. What's worse, PFedBA, a recent backdoor on PFL, shows that it can easily bypass most defenses. To address these limitations, we propose `EarlyShield`, an effective and data-free defense tailored for both FL and PFL. Our intuition is that even under PFL, benign clients exhibit multi-view consistency, while malicious updates tend to deviate in similarity structure and low-dimensional representations. `EarlyShield` leverages this idea while focusing on enforcing stringent early screening: (i) client screening based on similarity and principal component analysis (PCA), and (ii) similarity-driven decay to further suppress suspicious updates before aggregation. Extensive experiments on various datasets across independent and identically distributed (IID) and non-IID settings show that `EarlyShield` reduces attack success rates with minimal accuracy drop, consistently outperforming existing defenses. We open source the code as well.

Keywords: Personalized Federated Learning · Backdoors · Defenses

1 Introduction

Federated Learning (FL) enables multiple clients to collaboratively train a global model without sharing their private data. In each round, clients perform local training and send model updates to a central server, which aggregates them into a shared global model. While FL can provide strong privacy benefits,

its open nature exposes the global model to *low-cost backdoor attacks* launched through malicious client updates [1,2]. Personalized Federated Learning (PFL) extends FL by introducing client-specific components, such as personalized layers, to better accommodate heterogeneous data distributions [3]. However, such personalization of PFL also introduces **new security challenges** regarding defending against backdoors. Recent works [4,5] have shown that PFL is even more vulnerable to backdoor attacks when compared to FL. As shown in Fig. 1, benign model updates (i.e., blue crosses), particularly under personalization, tend to lead to a large volatility, which creates space for malicious model updates (i.e., red crosses) to appear as 'normal'.



(a) FL, IID (b) FL, non-IID (c) PFL, IID (d) PFL, non-IID

Fig. 1: An example to illustrate benign and malicious model updates, i.e., blue and red crosses, in representation space on CIFAR-10.

Although defenses have been developed to protect FL against backdoors, they become unreliable when directly applied to PFL due to the increased variability among benign updates themselves. In particular, clustering-based defenses assume that benign and malicious updates form clearly separable clusters, but this assumption no longer holds in PFL, where personalization naturally enlarges the spread of benign updates (see Fig. 1). Meanwhile, aggregation-based defenses still allow all clients, malicious ones included, to participate in every aggregation round, enabling gradual backdoor accumulation. In addition, some defenses further assume access to a small auxiliary dataset on the server side, an assumption that could be impractical. What's worse, **PFedBA** [5], a recent stealthy backdoor on PFL, shows that it could bypass most existing defenses. Our experiments later confirmed this conclusion as well.

In this paper, we primarily aim to answer two questions. That is, "**Q1:** Why do earlier strong defenses fail against strong backdoor attacks on PFL, such as PFedBA" and "**Q2:** How can we improve existing defenses on FL so that they become more effective on PFL?" As a result, we propose `EarlyShield`, a lightweight yet effective defense framework tailored to the characteristics of PFL while remaining compatible with classical FL. We derived EarlyShield mainly from revisiting earlier defenses. *We found that a stringent early stage of screening clients' model updates was critical, particularly for PFL.* Such a screening stage helps enhance robustness of FL as well. While the finding seems simple and intuitive, unfortunately, it has been overlooked in earlier works. In effect, EarlyShield decouples the defense into two main stages: (i) an early-stage screening procedure that selects potentially benign clients using a dual-view criterion, and (ii)

a similarity-driven decay mechanism that suppresses anomalous updates during aggregation. We emphasize that unlike previous defenses, EarlyShield does not require any auxiliary server-side data, does not assume clustering separability, and does not rely solely on weighting without filtering.

We further summarize our contributions as follows.

- We propose EarlyShield, a new defense mechanism designed specifically for PFL and compatible with FL and against recent strong backdoors.
- We identify the importance of an early stage of screening procedure from revisiting the general defense pipeline against backdoor attacks.
- We conduct extensive experiments on three datasets under both PFL and FL settings. Our results show that EarlyShield consistently reduces attack success rates while preserving high accuracy. We further perform a comprehensive ablation study. The code is at https://github.com/Shixiong-Li/PFedBA.

2 Related Work

2.1 Backdoor attacks and defenses in FL.

Backdoor attacks in FL. FL has witnessed various model poisoning attacks (MPAs), which can be broadly categorized as untargeted or targeted. Untargeted MPAs aim to degrade global model performance, e.g., by crafting malicious updates that are opposite to benign ones (e.g., *Krum-Attack* [6]). In the meanwhile, targeted backdoor attacks aim to force model predictions to be a targeted label by crafting and inserting backdoor patterns into samples. Well-known examples include *BadNets* [7], *DBA* [8], and others [1,9,10,11].

Backdoor defenses in FL. In general, many defenses have been proposed for FL. Specifically, there are aggregation-based methods such as *Trimmed Mean* [12], *FLTrust* [13], and *FLARE* [14] that assign weights to model updates of different clients based on metrics such as similarity score and apply weighted aggregation for updating the global model. There are also clustering-based approaches like *FLAME* [15], *GeminiGuard* [16], and others [17] that identify and filter suspicious updates.

2.2 Backdoor attacks and defenses in PFL.

Backdoor attacks in PFL. Unlike conventional FL, PFL seeks to optimize a collection of client-specific models that better adapt to their local data distributions. Recent literature reveals that attackers can exploit personalization strategies for their own purposes, i.e., to design more stealthy and persistent backdoors. Ye et al. [4] first proposed *BapFL*, which selectively poisons the global model while maintaining a diverse set of classifiers across clients. Later on, Lyu et al. [5] proposed *PFedBA*, which improved attack stealthiness by embedding the trigger pattern into features that are co-optimized with local training. The authors confirmed that PFedBA achieved high attack success rates against most defenses, thus raising serious concerns.

Backdoor defenses in PFL. Fan et al. [18] proposed *FLIGHT*, a lightweight secure aggregation method for PFL that leverages client clustering to restrict malicious updates. Another defense direction focuses on *robust parameter shar-*

ing, where only a smaller subset of model parameters is allowed to be aggregated such as in RobustPFL [19].

2.3 Uniqueness of EarlyShield.

EarlyShield is unique in that it primarily focuses on early-stage screening, which was previously neglected. We design EarlyShield to be model agnostic, overall lightweight, and effective against PFedBA, one of the strongest backdoors on PFL. EarlyShield also shows better performance when applied to FL and compared to baselines.

3 System and Adversary Model

System Model. In FL/PFL, a central parameter server (PS) coordinates training among N distributed clients $\{u_1, u_2, \ldots, u_N\}$, each holding private data that never leaves the local device. The PS maintains a global model $G_t \in \mathbb{R}^d$ at communication round t, and each client u_i maintains its own local model parameters $\theta_i \in \mathbb{R}^d$ within the same parameter space. During each round, the training process proceeds as below:

(1) The PS first sends the current G_t to the selected clients.

(2) Each participating client u_i initializes its local model as $\theta_i = G_t$ and performs several epochs of local training on its private dataset D_i, resulting in a model update $\Delta_i = \delta_i - G_t$. **For FL**, all clients share the same model architecture and loss function, whereas **for PFL**, the local training additionally includes client-specific components to capture user-level heterogeneity. Regardless of FL or PFL, model updates are then uploaded to the PS for model aggregation.

(3) Upon receiving all local updates, the PS applies various aggregation methods such as FedAvg-FT [20] for updating G_t. The process repeats for T communication rounds until convergence.

Adversary Model. Similar to earlier works, we consider two goals for backdoor attackers. That is, (1) attackers aim to achieve high ASRs on backdoor samples while (2) ensuring that the victim model, i.e., G_t, achieves high accuracy on clean samples.

As in earlier works, we assume that the attackers are legitimate participants in FL/PFL. More critically, they can manipulate their local data, e.g., by injecting trigger patterns and modifying labels, and craft the model updates submitted to PS to achieve the two forementioned goals. As participants, the attackers have white-box access to G_t but no access to local models of other clients. We assume the number of attackers is less than half of the total client population. PS is trusted and deploys defense mechanisms such as EarlyShield to mitigate potential backdoor attacks.

4 Design of EarlyShield

4.1 Overview

Here we introduce how PS deploys EarlyShield in its FL/PFL pipeline to effectively defend against backdoor attacks, particularly recent strong ones like PFedBA. EarlyShield works in a typical two-stage process, which is shown in

Fig. 2. Upon receiving model updates from selected clients, PS is to first apply stringent screening in order to largely prevent suspicious ones from participating in the following aggregation stage. In particular, we incorporate a dual-view selection criteria based on both cosine similarity and PCA, ensuring that the selected model updates from this stage exhibit sufficient consistency under both spatial and structural views. In the second stage, we propose a naive decay mechanism, i.e., re-using cosine similarity scores from earlier as a decay factor for robust aggregation of G_t. Our ablation studies confirm both stages are effective in contributing to defending against even recent backdoor attacks.

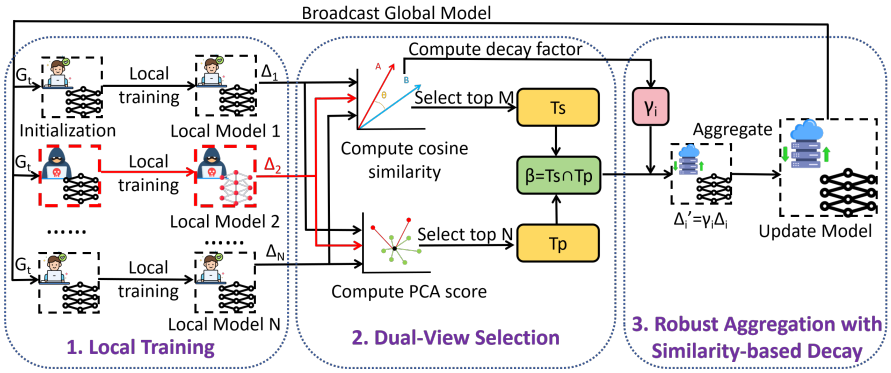


Fig. 2: The workflow of our proposed defense.

4.2 Core Components

Dual-View Selection As introduced above, EarlyShield uses dual-view selection criteria as the first screening stage. Denote the model update from Client $i \in N$ by $\Delta_i \in \mathbb{R}^d$. We first compute the pairwise cosine similarity between all client updates:

$$S_{ij} = \frac{\Delta_i \cdot \Delta_j}{\|\Delta_i\|_2 \, \|\Delta_j\|_2}, \quad \forall i, j \in \{1, \ldots, N\}. \tag{1}$$

For Client i, its average similarity with respect to all others can be computed as $\bar{S}_i = \frac{1}{N-1} \sum_{j \neq i} S_{ij}$. Therefore, our first selection criterion based on cosine similarity is that a higher $\bar{S}_i$ indicates that Δ_i complies better with the joint behavior of all model updates, hence suggesting benign intent [13,15].

We then compute the PCA score for each Δ_i, which reflects their structural consistency. Specifically, PS stacks all Δ_is into a matrix $\mathbf{U} \in \mathbb{R}^{N \times d}$ and projects it into the 2-dimensional space using PCA by $\mathbf{Z} = \mathrm{PCA}_2(\mathbf{U})$. We compute the Euclidean distance from each Δ_i's projection $\mathbf{z}_i$ to the global center $\bar{\mathbf{z}}$ as $D_i = \|\mathbf{z}_i - \bar{\mathbf{z}}\|_2$. By now, we can obtain the PCA-based score of Δ_i as $P_i = 1 - \frac{D_i}{\max_j D_j}$. Our second selection criterion is that a higher P_i indicates that Δ_i is closer to the structural center of all model updates, hence suggesting benign intent.

Combining two views together. We anticipate that the two above views could be complementary to each other and therefore proceed to select Δ_is with both high similarity and high PCA scores. Let $\mathcal{T}_S$ and $\mathcal{T}_P$ denote the top-M

and top-N clients ranked by $\bar{S}_i$ and P_i, respectively. PS proceeds to select the following set of clients for the next stage: $\mathcal{B} = \mathcal{T}_S \cap \mathcal{T}_P$. Note that such a combining strategy provides flexibility in selecting sufficient clients that have high $\bar{S}_i$ and P_i, which reduces the degradation of G_t.

Robust Aggregation with Similarity-based Decay In the aggregation stage of G_t, we propose incorporating a similarity-based decay mechanism to further condition the selected model updates from the previous stage. The main motivation is that some malicious model updates could inevitably be selected for the final aggregation, which can eventually sabotage G_t. The similarity-based decay mechanism is straightforward; we reward model updates that have a high average cosine similarity score with other model updates and vise versa.

In particular, we first compute the anomaly score (denoted by δ_i) for Δ_i as $\delta_i = 1 - \bar{S}_i$, where $\bar{S}_i$ is the mean cosine similarity of Δ_i with all other client updates (see Section 4.2). Intuitively, the higher δ_i, the lower $\bar{S}_i$, indicates that Δ_i deviates from the majority and therefore may be malicious. Then we normalize all δ_is following $\hat{\delta}_i = \frac{\delta_i - \min_j \delta_j}{\max_j \delta_j - \min_j \delta_j + \epsilon}$, where ϵ is a small constant to ensure that the denominator does not equal zero. Finally, we compute a weight factor for Δ_i as $\gamma_i = 1 - \alpha \cdot \hat{\delta}_i$, where α is the maximum decay. As a result, robust aggregation of EarlyShield works as: $\mathbf{G}_{t+1} = \mathbf{G}_t + \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \gamma_i \cdot \Delta_i,$, where $\mathcal{B}$ is the set of clients selected from the previous stage.

5 Performance Evaluation

5.1 Experimental Setup

Datasets and implementation. We evaluated our defense on three widely-used datasets: Fashion-MNIST, CIFAR-10, and CIFAR-100. These datasets are commonly adopted in the evaluation of both backdoor attacks and defenses. We adopted two data distribution settings as well: IID and non-IID. The non-IID setting was simulated using a Dirichlet distribution with its concentration parameter set to 0.5 as in [21]. We implemented EarlyShield using PyTorch.

FL and PFL setting. For all FL and PFL experiments, we assume a total of 100 clients with a sampling rate of 10% per round. The local batch size was 64, and the portion of malicious clients was 10%. For FL, each selected client ran 2 local epochs per round. Learning rates were 0.01 (Fashion-MNIST) and 0.1 (CIFAR-10/100). We used CNN for Fashion-MNIST and ResNet for CIFAR-10/100. The default poisoning ratio was 50%. Similarly for PFL, each clients ran 20 local epochs, followed by 1 (Fashion-MNIST) or 5 (CIFAR-10/100) personalized epochs. Learning rate was 0.1, with 25% data poisoning ratio. For all attacks, attackers' targeted label is class 0.

Evaluation metrics. In the experiments, we use the following two metrics for evaluating defenses: *(1) Model accuracy (ACC)*, i.e., the prediction accuracy of the victim model, *(2) average Attack success rate (average ASR)* [22], i.e., the proportion of accumulated backdoor samples misclassified into the target class over the number of epoches. We refer to it as ASR for convenience purpose.

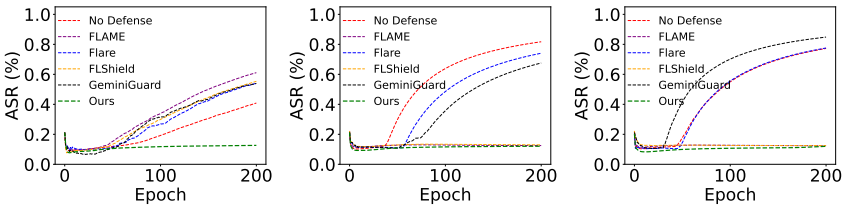
Defense baselines and backdoor attacks. We evaluated EarlyShield against four baselines: FLAME [15], Flare [14], FLShield [23] and GeminiGuard [16]. For

attacks on PFL, we adopt the most recent strong backdoor attack, PFedBA [5], across three different PFL algorithms: pFedMe [3], FedAvg-FT [20], and FedProx-FT [24]. For FL, we adopted the following three popular ones: BadNets [7], DBA [8], and Scaling Attack [1].

5.2 Defense Effectiveness of EarlyShield

Defense Robustness against Backdoor Attacks over PFL We summarize ASR and ACC of our defense and baselines across various datasets in Tables 1 to 3. Note that the lower ASR, the better defense. On Fashion-MNIST, our method achieves the lowest average ASR across all PFL methods and data distributions, while other defenses degrade severely in personalized scenarios like pFedMe. Notably, we observe that some defenses even result in higher ASRs compared to the no-defense scenario. We anticipate that the adopted defenses may incorrectly exclude benign model updates or assign lower weights to them, therefore amplifying the impact of malicious ones. We believe this is a unique challenge in PFL, where the inherent data heterogeneity makes it extremely difficult to distinguish benign clients from malicious ones. On CIFAR-10 and CIFAR-100, while there is no defense that can effectively tackle attacks, our approach still provides the most substantial ASR reduction in most cases, confirming its general effectiveness. We plot the average ASRs of different datasets but only show results on Fashion-MNIST under IID here **due to space limit**, as shown in Fig. 3. Results on other datasets including non-IIDs confirm the effectiveness of our method as well.

Defense Robustness against Backdoor Attacks over FL Here we plan to confirm whether EarlyShield performs better than other baselines on FL. Our motivation is to explore the importance of early-stage screening of EarlyShield in defending against backdoors, which other baselines did not have. *Due to space limit, we only report results on Fashion-MNIST.* However, consistent results were obtained on CIFAR-10 and CIFAR-100. As can be seen in Table 4, while some baselines achieve competitive performance in specific settings, they exhibit critical failures in other settings, particularly against DBA under non-IID, where ASR surges to nearly 100%. In contrast, EarlyShield achieves stable, low ASRs across all datasets and attack types without major failures. The results suggest that early-stage screening greatly enhances defense performance on FL.



(a) pFedMe, IID (b) FedAvg, IID (c) FedProx, IID

Fig. 3: Defense Robustness under PFedBA [5] (F-MNIST)

Table 1: Defense Robustness against Backdoor Attacks on PFL (F-MNIST)

Data Distribution →	IID						non-IID					
PFL Method →	pFedMe		FedAvg-FT		FedProx-FT		pFedMe		FedAvg-FT		FedProx-FT	
Metrics→	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
W/O Defense	40.87	88.54	81.72	90.56	77.26	90.92	34.89	86.75	78.24	89.32	73.93	86.69
FLAME	61.11	88.11	12.49	90.32	12.40	90.27	49.42	86.39	11.84	89.01	12.14	89.41
Flare	54.12	82.38	74.07	90.67	77.57	90.67	38.30	85.42	68.15	86.68	73.47	87.42
FLShield	55.33	88.41	12.97	90.42	12.49	90.91	51.90	86.79	11.57	89.38	12.70	89.55
GeminiGuard	53.97	83.48	67.57	90.51	84.85	89.61	55.30	83.24	52.57	86.23	78.49	84.27
Ours	**12.62**	86.87	**12.05**	88.86	**11.98**	88.80	**12.59**	84.33	**11.28**	87.26	**11.89**	87.42

Table 2: Defense Robustness against Backdoor Attacks on PFL (CIFAR-10)

Data Distribution →	IID						non-IID					
PFL Method →	pFedMe		FedAvg-FT		FedProx-FT		pFedMe		FedAvg-FT		FedProx-FT	
Metrics→	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
W/O Defense	84.76	60.66	89.57	74.29	88.64	73.09	75.59	44.07	90.12	62.19	87.52	61.76
FLAME	88.53	52.03	86.94	58.81	84.44	59.68	84.33	41.52	78.98	41.28	79.17	41.77
Flare	64.22	58.15	82.42	78.87	83.84	79.58	60.52	37.51	81.75	48.13	71.62	59.62
FLShield	6.93	54.11	85.63	68.87	81.01	65.37	71.03	37.67	80.48	40.44	85.40	42.95
GeminiGuard	83.67	69.63	86.26	79.89	85.09	79.05	52.47	26.11	83.39	46.04	76.08	42.31
Ours	**6.84**	58.83	**25.30**	64.03	**48.18**	63.90	**46.52**	45.27	**71.43**	45.23	**70.96**	45.72

Table 3: Defense Robustness against Backdoor Attacks on PFL (CIFAR-100)

Data Distribution →	IID						non-IID					
PFL Method →	pFedMe		FedAvg-FT		FedProx-FT		pFedMe		FedAvg-FT		FedProx-FT	
Metrics→	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
W/O Defense	70.01	15.94	90.54	37.25	88.22	38.05	69.15	13.20	91.46	34.53	88.69	35.28
FLAME	82.31	13.28	90.60	27.03	88.85	27.82	80.16	10.71	75.06	22.69	69.55	22.90
Flare	65.02	15.85	84.12	42.39	82.67	44.34	66.40	12.10	82.14	38.48	81.34	26.83
FLShield	10.33	15.75	83.35	29.22	79.66	29.11	45.47	17.43	84.92	26.02	80.31	41.22
GeminiGuard	75.78	19.14	84.65	44.55	83.50	44.97	58.05	13.30	84.18	40.35	82.29	41.65
Ours	**5.41**	12.57	**34.76**	27.11	**69.99**	29.13	**15.44**	9.88	**47.22**	24.84	**61.98**	21.33

Table 4: Defense Robustness under Attacks for FL (Fashion-MNIST)

Data Distribution →	IID						non-IID					
Attacks →	Badnets		DBA		Scaling		Badnets		DBA		Scaling	
Metrics→	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
W/O Defense	99.99	89.39	99.92	89.04	99.84	89.72	99.97	88.02	99.94	87.75	99.87	87.56
FLAME	**0.51**	88.86	**0.73**	88.58	0.62	88.99	1.57	85.13	99.98	85.08	2.33	85.39
Flare	1.38	88.93	0.80	85.16	2.40	88.89	1.35	87.35	100.00	87.12	1.32	86.82
FLShield	0.91	89.04	1.16	89.31	0.49	88.98	**0.75**	86.86	99.94	89.27	4.53	86.90
GeminiGuard	1.01	88.57	38.84	88.32	14.02	87.33	3.86	84.90	8.52	85.65	32.92	81.65
Ours	2.49	86.33	**0.73**	85.13	**0.47**	85.31	1.33	83.56	**7.20**	81.81	**1.27**	83.17

6 Ablation Study

6.1 Effectiveness of Individual Components

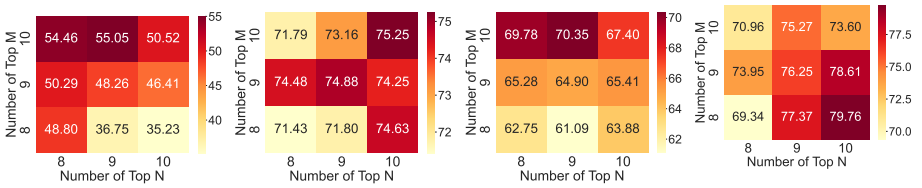
From Table 5, we can see that combining all components of our defense substantially reduces the ASR, indicating that each component contributes to mitigating backdoors. When removing any individual component, the ASR generally increases, further confirming the necessity of all three modules.

Table 5: Effectiveness of Individual Components on CIFAR-10

Data Distribution →	IID				non-IID			
PFL Method →	FedAvg-FT		FedProx-FT		FedAvg-FT		FedProx-FT	
Metrics→	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
W/O Defense	89.57	74.29	88.64	73.09	90.12	62.19	87.52	61.76
W/ All Components	48.58	63.67	62.75	63.70	71.43	45.23	70.96	45.72
W/O Similarity-Based Selection	54.46	62.58	69.78	63.93	71.79	43.42	69.34	43.12
W/O PCA-Based Selection	35.23	66.53	63.88	66.78	74.63	46.13	73.60	43.39
W/O Similarity-Based Decay	68.80	83.68	71.07	66.37	83.68	50.92	74.06	44.68

6.2 Impacts of M and N in Dual-View Selection.

Fig. 4 reports the effect of varying the number of selected clients, i.e., M and N, in the dual-view selection stage under both IID and non-IID settings using FedAvg-FT and FedProx-FT. Under **IID**, a clear pattern emerges. Increasing Top-M consistently leads to higher ASR, which is expected given that admitting more clients based on similarity weakens the ability to exclude malicious updates. In contrast, increasing Top-N generally reduces ASR for both FedAvg-FT and FedProx-FT. This indicates that PCA-based selection may be aggressively filtering clients and, in some cases, discarding benign ones. Based on the results, $M = 10$ and $N = 8$ could be a good balance when PS receives 10 model updates. In the **non-IID** case, the trend is less pronounced. We anticipate that this may be due to data heterogeneity. To avoid over-pruning benign clients, we adopt a more conservative choice of $M = 8$ and $N = 8$.



(a) FedAvg-FT,IID (b) FedAvg-FT,non-IID (c) FedProx-FT,IID (d) FedProx-FT,non-IID

Fig. 4: ASR under different M and N on CIFAR-10

6.3 Impact of α in Similarity-based Decay

Table 6 reports the impact of different α on both ASR and ACC. Overall, increasing α consistently reduces ASR in all settings, demonstrating that stronger decay suppresses malicious client updates more effectively. This effect is especially pronounced under the IID setting, where cosine similarity serves better to indicate how well a model update complies with the majority. As expected, a larger decay factor also leads to a reduction in ACC, thus raising the need of a better trade-off between robustness and model utility.

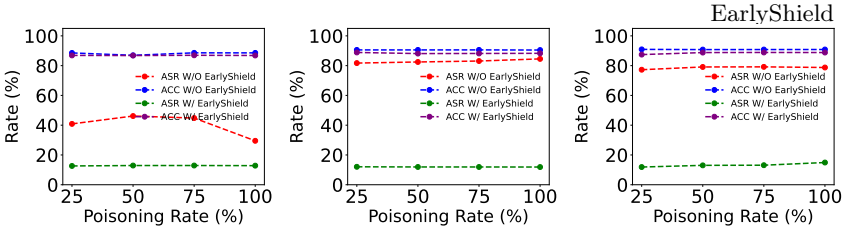
To this end, we select different decay strengths for IID and non-IID settings. Under IID data, $\alpha = 1.0$ achieves the best ASR reduction with an acceptable accuracy drop. Under the more heterogeneous non-IID distribution, we adopt a more conservative choice of $\alpha = 0.5$, which preserves accuracy while still providing effective suppression of backdoor behaviors.

Table 6: Influence of Decay Strength (α) on CIFAR-10

Data Distribution →	IID				non-IID			
PFL Method →	FedAvg-FT		FedProx-FT		FedAvg-FT		FedProx-FT	
Metrics→	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
$\alpha = 0.0$	68.93	68.80	71.07	66.37	83.68	50.92	74.06	44.68
$\alpha = 0.3$	44.06	67.27	68.34	68.82	73.09	46.42	70.59	44.72
$\alpha = 0.5$	35.23	66.53	62.75	65.79	71.43	45.23	70.96	45.72
$\alpha = 0.7$	26.73	65.73	57.92	65.70	70.10	45.47	68.32	42.97
$\alpha = 1.0$	25.30	64.03	51.60	64.66	64.32	45.16	65.36	42.88

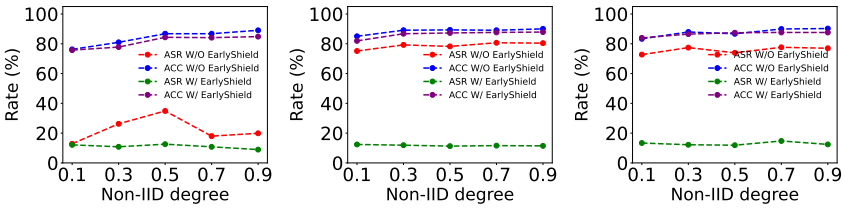
6.4 Impact of Data Poisoning Rate.

We evaluate how different poisoning rates impact ASR and ACC under both IID and non-IID settings, which show consisting results. Due to space limit, we only plot results under IID as in Fig. 5. We can see that, without any defense, the ASR increases as the poisoning rate increases. After applying our defense, however, the ASR remains consistently low across all poisoning rates and data distributions, demonstrating the robustness of our method towards various attack strengths.



(a) pFedMe, IID (b) FedAvg-FT, IID (c) FedProx-FT, IID

Fig. 5: Impact of Data Poisoning Rate (Fashion-MNIST)



(a) pFedMe (b) FedAvg-FT (c) FedProx-FT

Fig. 6: Impact of non-IID Degree (Fashion-MNIST)

6.5 Impact of Non-IID Degree.

We further study how different levels of data heterogeneity, i.e., non-IID degrees, affect the robustness of EarlyShield in Fig.6. Recall that we simulate non-IID using a Dirichlet distribution as in [21]. Fig.6 shows that across all non-IID degrees, our defense maintains consistently strong robustness.

7 Conclusion

In this paper, we presented `EarlyShield`, a practical and data-free defense designed for both PFL and FL. Our experiments showed that existing defenses become unreliable in PFL. By focusing on early-stage screening, `EarlyShield` was able to exploit a simple yet effective defense strategy against various backdoor attacks on PFL. Extensive experiments across IID and non-IID confirm that `EarlyShield` substantially reduces ASR while maintaining accuracy.

Acknowledgments. We would like to thank anonymous reviewers for their constructive comments and helpful advice.

References

1. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: AISTATS. pp. 2938–2948. PMLR (2020)
2. Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.y., Lee, K., Papailiopoulos, D.: Attack of the tails: Yes, you really can backdoor federated learning. NeurIPS **33**, 16070–16084 (2020)
3. T Dinh, C., Tran, N., Nguyen, J.: Personalized federated learning with moreau envelopes. NeurIPS (2020)
4. Ye, T., Chen, C., Wang, Y., Li, X., Gao, M.: Bapfl: You can backdoor personalized federated learning. TKDD (2024)

5. Lyu, X., Han, Y., Wang, W., Liu, J., Zhu, Y., Xu, G., Liu, J., Zhang, X.: Lurking in the shadows: Unveiling stealthy backdoor attacks against personalized federated learning. In: USENIX Security (2024)
6. Fang, M., Cao, X., Jia, J., Gong, N.: Local model poisoning attacks to {Byzantine-Robust} federated learning. In: USENIX Security (2020)
7. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019)
8. Xie, C., Huang, K., Chen, P.Y., Li, B.: Dba: Distributed backdoor attacks against federated learning. In: ICLR (2019)
9. Bi, Z., Singha, A., Xue, H., Li, T., Chen, Y., Zhang, Y.: Physical backdoor attacks against mmwave-based human activity recognition. In: IEEE ICDCS. pp. 758–768. IEEE (2025)
10. Li, S., Lyu, X., Wang, N., Li, T., Chen, D., Chen, Y.: Beyond uniformity: Robust backdoor attacks on deep neural networks with trigger selection. In: PAKDD. pp. 290–302. Springer (2025)
11. Jiang, Z., Lyu, X., Shi, S., Xiao, Y., Chen, Y., Hou, Y.T., Lou, W., Wang, N.: Boba: Boosting backdoor detection through data distribution inference in federated learning. In: ECAI. Frontiers in Artificial Intelligence and Applications, vol. 413, pp. 1051–1058 (2025). https://doi.org/10.3233/FAIA250914, research Article
12. Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-robust distributed learning: Towards optimal statistical rates. In: ICML (2018)
13. Cao, X., Fang, M., Liu, J., Gong, N.Z.: Fltrust: Byzantine-robust federated learning via trust bootstrapping. In: NDSS (2021)
14. Wang, N., Zhang, C., Xiao, Y., Chen, Y., Lou, W., Hou, Y.T.: Flare: Defending federated learning against model poisoning attacks via latent space representations. IEEE TDSC (01), 1–17 (2024)
15. Nguyen, T.D., Rieger, P., Chen, H., Yalame, H., Möllering, H., Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Zeitouni, S., et al.: {FLAME}: Taming backdoors in federated learning. In: USENIX Security (2022)
16. Lyu, X., Wang, N., Xiao, Y., Li, S., Li, T., Chen, D., Chen, Y.: Two heads are better than one: Model-weight and latent-space analysis for federated learning on non-iid data against poisoning attacks. arXiv preprint arXiv:2503.23288 (2025)
17. Lyu, X., Wang, N., Xiao, Y., Li, S., Li, T., Chen, D., Chen, Y.: Buffer is all you need: Defending federated learning against backdoor attacks under non-iids via buffering. In: IEEE TrustCom. pp. 236–243. IEEE (2025)
18. Fan, T., Chen, X., Dong, Y., Chen, X., Xuan, Y., Jing, W.: Lightweight secure aggregation for personalized federated learning with backdoor resistance. In: ACSAC (2024)
19. Chen, G., Wang, W., Wu, Y., Li, C., Xu, G., Ji, S., Li, T., Shen, M., Han, Y.: Robustpfl: Robust personalized federated learning. IEEE TDSC (2025)
20. Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., Ramage, D.: Federated evaluation of on-device personalization. arXiv preprint arXiv:1910.10252 (2019)
21. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. In: IEEE ICDE (2022)
22. Zhuang, H., Yu, M., Wang, H., Hua, Y., Li, J., Yuan, X.: Backdoor federated learning by poisoning backdoor-critical layers. In: ICLR (2024)
23. Kabir, E., Song, Z., Rashid, M.R.U., Mehnaz, S.: Flshield: a validation based federated learning framework to defend against poisoning attacks. In: IEEE S&P (2024)
24. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. MLSys (2020)