SD-CSFL: A Synthetic Data-Driven Conformity Scoring Framework for Robust Federated Learning –Appendix

Ebtisaam Alharbi[1,2] Abdulrahman Kerim[3] Leandro Soriano Marcolino[4,2] Qiang Ni[2]
easharbi@uqu.edu.sa a.kerim@surrey.ac.uk leandro.sm@vinuni.edu.vn q.ni@lancaster.ac.uk
[1]Umm Al-Qura University, Saudi Arabia [2]Lancaster University, United Kingdom
[3]University of Surrey, United Kingdom [4]VinUniversity, Vietnam

1. Synthetic Data Generation Methodology

Overview of Synthetic Data Generation. The construction of the calibration dataset $\mathcal{D}_{\text{calibration}}$ is a pivotal aspect of our approach. We employ a synthetic data generation pipeline inspired by prior work [6], integrating Stable Diffusion-V2 [10] and ChatGPT-3.5 [9]. This methodology ensures the creation of high-quality, domain-relevant data tailored to the specific requirements of federated learning tasks. By focusing on diversity and domain relevance, the synthetic dataset becomes well-suited for computing nonconformity scores while simultaneously preserving privacy and eliminating reliance on potentially compromised client data.

Attribute Extraction: Defining Task-Relevant Features. The synthetic data generation begins with attribute extraction, which identifies key characteristics relevant to the task domain. Attributes such as object types, colors, textures, poses, and semantic relationships are curated using ChatGPT-3.5 [9]. For example, in bird-related datasets, extracted attributes might include species (e.g., ”sparrow” or ”parrot”), feather colors, environmental contexts (e.g., ”perched on a branch under a cloudy sky”), and other descriptive details. This phase ensures the synthetic data aligns with the nuances and variability of real-world distributions, providing a meaningful basis for nonconformity score calculations.

Prompt Creation: Translating Attributes into Descriptions. The extracted attributes are systematically transformed into descriptive text prompts for Stable Diffusion. These prompts are carefully crafted to combine multiple attributes into meaningful, context-rich descriptions. For instance, a prompt like ”Generate an image of a parrot with green feathers sitting on a wooden branch surrounded by a forest canopy” incorporates attributes such as color, species, posture, and environment. Randomized sampling of attributes adds further variability to the prompts, ensuring that the generated dataset captures a broad range of potential real-world scenarios.

Image Generation: Creating Diverse and Realistic Data. The generated prompts are processed by Stable Diffusion-V2 [10], a state-of-the-art text-to-image generative model. Using the stochastic nature of the model and introducing controlled variations in prompt parameters, the synthetic data pipeline produces high-quality images that are both photorealistic and domain-relevant. This step ensures the calibration dataset exhibits sufficient diversity to capture inherent task variability, making it suitable for evaluating the nonconformity of model updates.

2. Attack Methods

We evaluated the effectiveness of our SD-CSFL method against several recent and sophisticated adversarial attacks in FL.

Adversarially Adaptive Backdoor Attack to Federated Learning (A3FL): A3FL enhances backdoor persistence by dynamically adapting the trigger to the global training dynamics [13]. It optimizes the trigger for both the current and adversarially crafted global models, ensuring its effectiveness through multiple updates. By utilizing adversarial adaptation loss and Projected Gradient Descent (PGD), A3FL continuously refines the backdoor, allowing it to survive the evolving training process.

Focused-Flip Federated Backdoor Attack (F3BA): F3BA targets a small subset of model parameters, altering them with minimal impact on overall performance [4]. The importance of each parameter, $S[j]$, is evaluated as: $S[j] = -\left(\frac{\partial L_g}{\partial w[j]}\right) \odot w[j]$. The selected parameters' signs are flipped to align with a trigger pattern, embedding the backdoor without causing significant deviation from normal model behavior.

Cerberus Poisoning Backdoor Attack (CerP): CerP introduces a stealthy and distributed backdoor attack by fine-tuning backdoor triggers, controlling local model parameter biases, and maximizing diversity among malicious updates. By exploiting defense assumptions, CerP minimizes deviations between poisoned and benign models, achieving high attack success rates while preserving the main learning task's accuracy [8].

Inner Product Manipulation Attack (IPM). The IPM attack aims to evade detection by aligning the malicious gradient with the direction of the true gradient while maintaining the same norm. This manipulation degrades the model's performance without raising suspicion. The attacker crafts a malicious update vector $\Delta \mathbf{g}_t^i$ using:

$$\Delta \mathbf{g}_t^i = \epsilon \cdot \text{sign}(\mathbf{g}_t^i \odot \mathbf{w}),$$

where $\mathbf{g}_t^i$ is the gradient of client i at iteration t, $\mathbf{w}$ is the model parameter vector, $\odot$ denotes element-wise multiplication, and ϵ is a scalar controlling the attack strength [12]. This formulation ensures the malicious update maintains inner product alignment with benign updates, making detection more difficult.

A Little is Enough Attack (ALIE). The ALIE attack perturbs updates within a plausible statistical range to bypass anomaly detectors. Assuming that benign gradients follow a typical distribution, the attacker estimates the coordinate-wise mean $\boldsymbol{\mu}_j$ and standard deviation $\boldsymbol{\sigma}_j$ from benign clients, and then sets:

$$\Delta \mathbf{g}_i^{(j)} \in \left[\boldsymbol{\mu}_j - z_{\max} \cdot \boldsymbol{\sigma}_j,\ \boldsymbol{\mu}_j + z_{\max} \cdot \boldsymbol{\sigma}_j\right],$$

for each coordinate $j \in \{1, \ldots, d\}$, where $z_{\max}$ is a tunable threshold (e.g., derived from the cumulative standard normal distribution) [1]. This allows the malicious updates to mimic benign variability while disrupting model convergence.

3. Datasets and Models.

We conduct experiments on the CIFAR-10 and Birds datasets [5, 7], along with their synthetic counterparts, CIFAR-10-Synth and Birds-Synth [6]. In our setup, real datasets are used for local training on client devices, while the synthetic datasets serve exclusively as calibration data for computing nonconformity scores on the server side. This separation ensures that the calibration process remains privacy-preserving and independent of potentially compromised client data.

For CIFAR-10, we employ a CNN with three convolutional layers (with 32, 64, and 128 filters), followed by batch normalization, ReLU activation, MaxPooling, and a fully connected layer with 256 units, ReLU activation, and 0.25 dropout. The output layer consists of 10 classes. Training uses a batch size of 64, a learning rate of 0.01, and $50,000$ training samples [3], with server calibration performed on $14,523$ CIFAR-10-Synth samples.

The Birds dataset [11], comprising 525 fine-grained species, presents a challenging classification task due to high inter-class similarity, significant intra-class variability, and substantial class imbalance. We use a ResNet50 model, pre-trained and fine-tuned on `layer4`, with a 1024-unit fully connected (FC) classifier, batch normalization, ReLU activation, 0.5 dropout, and an FC layer for 525 classes. Training uses a batch size of 16, a learning rate of 0.001, and 84,635 training samples, with server calibration performed on 20,475 Birds-Synth samples.

4. Additional Experiments

No Attack Model. We evaluated the performance of FL aggregator methods in a clean environment without attacks under Non-IID conditions. The results demonstrate that, in the absence of attacks, the attack success rate remains close to zero across all methods, as shown in Figure 1. This indicates robust model behavior in scenarios where no adversarial model occurs. Most methods also exhibited good accuracy, except the Krum [2] model, which, as shown in Figure 1, demonstrated reduced accuracy due to its limitation of selecting only one model as the aggregator, making it less effective under Non-IID conditions.

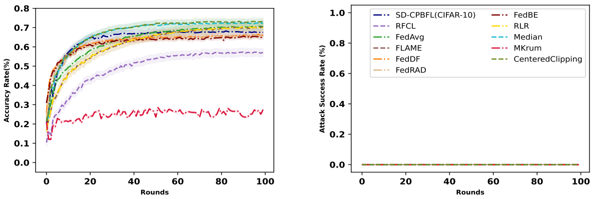
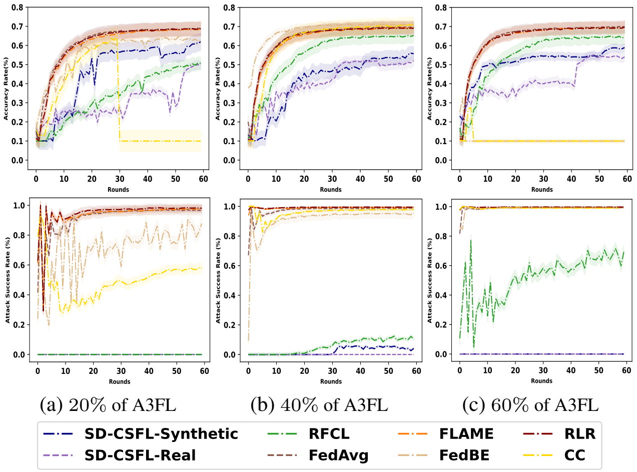


Figure 1. Performance of Baselines and SD-CSFL on CIFAR-10 Under Non-IID ($\alpha = 0.9$) Against No Attack

Effective Defense Against A3FL Attacks.

The experimental results, as shown in Figure 2, demonstrate the SD-CSFL framework's robustness in defending against A3FL [13] attacks under challenging Non-IID conditions ($\alpha = 0.5$). The analysis of accuracy rates and attack success rates across different attack intensities (20%, 40%, and 60%) provides a clear assessment of the framework's effectiveness.

Across the 20%, 40%, and 60% A3FL attack scenarios, the SD-CSFL framework consistently maintains stable accuracy rates, even as attack intensity increases. For the 20% A3FL attack, accuracy remains steady around 0.6, indicating strong resilience to moderate adversarial interference. At 40% A3FL attack intensity, accuracy fluctuates slightly but remains above 0.5, showing the framework's ability to preserve model performance under more intense attacks. Even at 60% A3FL attack intensity, accuracy stays above 0.4, highlighting the framework's robustness in Non-



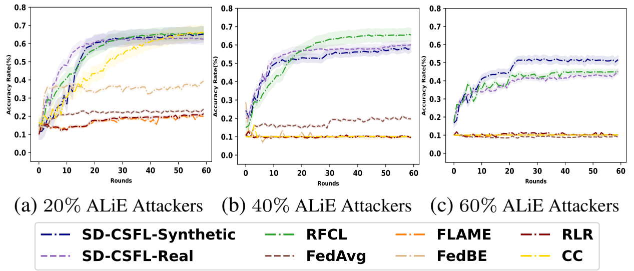
(a) 20% of A3FL (b) 40% of A3FL (c) 60% of A3FL

Figure 2. Performance of baselines and SD-CSFL on CIFAR-10 against A3FL attack under Non-IID ($\alpha = 0.5$).

IID conditions.

The attack success rates further confirm the framework's effectiveness. As shown, under a 20% A3FL attack, the success rate is low, staying below 0.2, suggesting that most adversarial updates are effectively neutralized. At 40% A3FL attack intensity, the success rate rises moderately, reaching up to 0.4, but remains controlled, reflecting the framework's continued ability to mitigate adversarial influence. Although the success rate peaks around 0.6 under a 60% A3FL attack, the framework still prevents a full compromise, demonstrating its resilience against severe attacks.

These results underscore the SD-CSFL framework's effectiveness in Non-IID environments. The adaptive percentile thresholds and balanced calibration set, central to the framework, significantly enhance its ability to detect and mitigate malicious updates, even when faced with the complexities of Non-IID data distributions. In summary, the SD-CSFL framework provides a robust defense against A3FL attacks, effectively protecting the global model in challenging Non-IID conditions and outperforming existing defenses.



(a) 20% ALiE Attackers (b) 40% ALiE Attackers (c) 60% ALiE Attackers

Figure 3. Performance of baselines and SD-CSFL on CIFAR-10 against ALiE attack under Non-IID ($\alpha = 0.5$).

Effective Defense Against ALiE Attacks.

The experimental results, as shown in Figure 3, demonstrate the effectiveness of the SD-CSFL framework in defending against ALiE attacks under Non-IID conditions ($\alpha = 0.5$). Despite the increasing intensity of the attacks—20%, 40%, and 60%—the framework consistently exhibits strong resilience.

In a challenging Non-IID environment, the SD-CSFL framework maintains the integrity of the global model, even as the intensity of the ALiE attacks escalates. While some impact on model performance is observed, the framework's adaptive mechanisms, including dynamic thresholds and a balanced calibration set, effectively mitigate these adverse effects. This resilience is particularly significant given the subtle nature of ALiE attacks, which aim to manipulate the model without easy detection.

The consistent performance across all scenarios underscores the robustness of the SD-CSFL framework and its effectiveness in countering sophisticated adversarial strategies. These results justify the adoption of SD-CSFL in federated learning settings where data heterogeneity and adversarial threats are prevalent.

Effective Defense Against A3FL and F3BA Attacks.

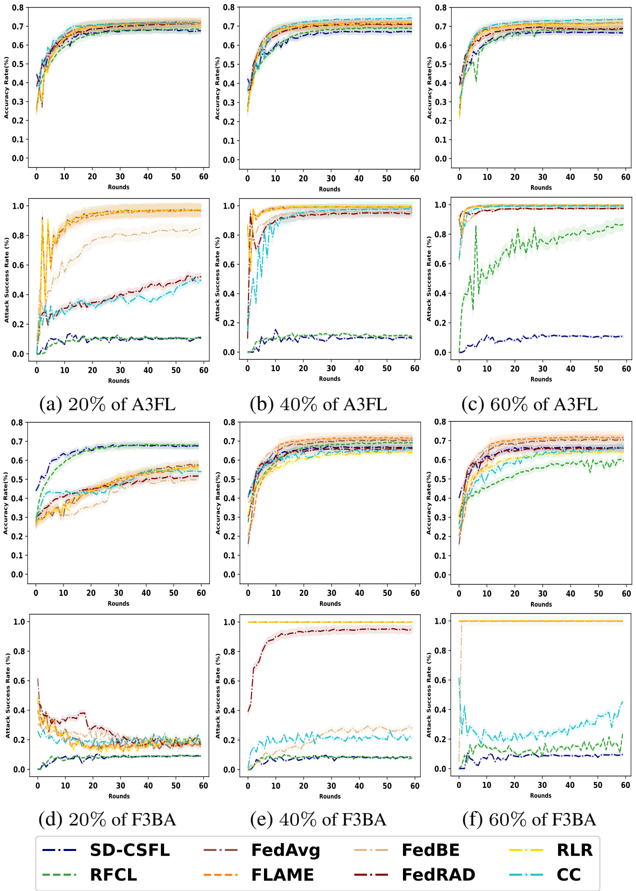
The experimental results in Figure 4 provide a clear demonstration of the SD-CSFL framework's robust defense against both A3FL and F3BA attacks under IID conditions. Despite varying attack intensities—20%, 40%, and 60%—the framework consistently maintains its effectiveness with minimal performance degradation.

Under A3FL and F3BA attacks, the SD-CSFL framework effectively mitigates adversarial influence, ensuring that both accuracy rates and attack success rates remain within controlled bounds. The framework's adaptive mechanisms, including dynamic thresholds and balanced calibration sets, play a critical role in preserving model integrity. Even as the intensity of the attacks increases, the framework demonstrates strong resilience, preventing significant compromise of the model's performance.

These results collectively underscore the SD-CSFL framework's capability to protect federated learning models from sophisticated adversarial threats. Its consistent performance across different attack types and intensities justifies its adoption in environments where maintaining model accuracy and security is paramount.

Effective Defense Against ALiE Attacks. The Figure 5 demonstrates the SD-CSFL framework's robust defense against ALiE attacks under IID conditions. Despite varying attack intensities—20%, 40%, and 60%—the framework consistently maintains high accuracy rates, effectively neutralizing the adversarial impact. These findings confirm the framework's reliability in safeguarding federated learning models from sophisticated attacks.

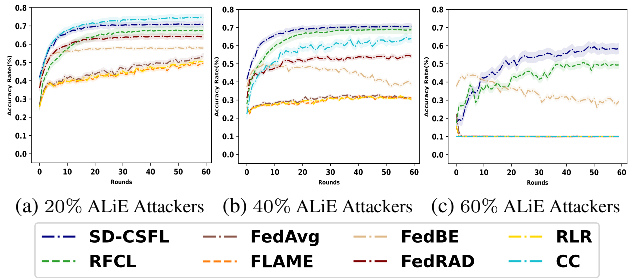
Additional Ablation Study under IID. The Figure 6 show



(a) 20% of A3FL (b) 40% of A3FL (c) 60% of A3FL

(d) 20% of F3BA (e) 40% of F3BA (f) 60% of F3BA

Figure 4. Performance of baselines and SD-CSFL on CIFAR-10 against A3FL and F3BA attacks under IID.



(a) 20% ALiE Attackers (b) 40% ALiE Attackers (c) 60% ALiE Attackers

Figure 5. Performance of baselines and SD-CSFL on CIFAR-10 against ALiE attack under IID.

that under IID conditions, there is no significant difference in the performance of the SD-CSFL framework when using a balanced versus a non-balanced calibration set. Both calibration approaches result in similar accuracy rates and attack success rates, indicating that the uniform distribution of data inherent in IID conditions mitigates the need for a balanced calibration set. This suggests that, under IID scenarios, the SD-CSFL framework's effectiveness is maintained regardless of the calibration set's balance, highlighting the framework's robustness in environments where data is evenly distributed.

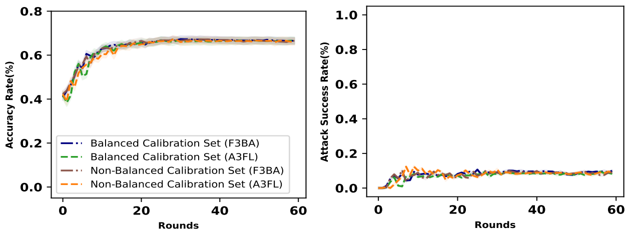


Figure 6. Impact of balanced vs. non-balanced calibration set under IID

The Impact of Selecting Percentile Thresholding. The choice of percentile thresholds in the SD-CSFL framework is crucial for accurately distinguishing between benign and malicious updates. Low thresholds increase the risk of misclassifying benign updates as malicious, while high thresholds may allow adversarial updates to evade detection. The experimental results highlight that optimal threshold selection is key to balancing detection accuracy and maintaining overall model performance, ensuring robust defense across various scenarios.

References

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

[2] 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. 2

[3] Rahul Chauhan, Kamal Kumar Ghanshala, and RC Joshi. Convolutional neural network (cnn) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 278–282. IEEE, 2018. 2

[4] Pei Fang and Jinghui Chen. On the vulnerability of backdoor defenses for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[5] Gpiosenka. Birds 525 species - image classification. https://www.kaggle.com/datasets/gpiosenka/100-bird-species, 2020. [online] Available: https://www.kaggle.com/datasets/gpiosenka/100-bird-species. 2

[6] Abdulrahman Kerim, Leandro Soriano Marcolino, Erickson R Nascimento, and Richard Jiang. Multi-armed bandit approach for optimizing training on synthetic data. *arXiv preprint arXiv:2412.05466*, 2024. 1, 2

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. *University of Toronto, Tech. Rep.*, 2009. 2

[8] Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Jiqiang Liu, and Xiangliang Zhang. Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9020–9028, 2023. 2

[9] OpenAI. ChatGPT. `https://chat.openai.com/`, 2024. Online; accessed: 2024-07-20. 1

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[11] Mangalam Sankupellay and Dmitry Konovalov. Bird call recognition using deep convolutional neural network, resnet-50. In *Proc. Acoustics*, pages 1–8, 2018. 2

[12] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pages 261–270. PMLR, 2020. 2

[13] Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3fl: Adversarially adaptive backdoor attacks to federated learning. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2