# FLGUARD: Secure and Private Federated Learning


Thien Duc Nguyen[1], Phillip Rieger[1], Hossein Yalame[2], Helen Möllering[2], Hossein Fereidooni[1], Samuel Marchal[3], Markus Miettinen[1], Azalia Mirhoseini[4], Ahmad-Reza Sadeghi[1], Thomas Schneider[2], and Shaza Zeitouni[1]

[1]System Security Lab, TU Darmstadt, Germany - {ducthien.nguyen, phillip.rieger, hossein.fereidooni, markus.miettinen, ahmad.sadeghi, shaza.zeitouni}@trust.tu-darmstadt.de

[2]Encrypto, TU Darmstadt, Germany - {yalame, moellering, schneider }@encrypto.cs.tu-darmstadt.de

[3]Aalto University and F-Secure, Finland - samuel.marchal@aalto.fi

[4]Google, USA - azalia@google.com



***Abstract*—Recently, a number of *backdoor attacks* against Federated Learning (FL) have been proposed. In such attacks, an adversary injects poisoned model updates into the federated model aggregation process with the goal of manipulating the aggregated model to provide false predictions on specific adversary-chosen inputs. A number of defenses have been proposed but none of them can effectively protect the FL process also against so-called multi-backdoor attacks in which multiple different backdoors are injected by the adversary simultaneously without severely impacting the benign performance of the aggregated model. To overcome this challenge, we introduce FLGUARD, a poisoning defense framework that is able to defend FL against state-of-the-art backdoor attacks while simultaneously maintaining the benign performance of the aggregated model. Moreover, FL is also vulnerable to *inference attacks*, in which a malicious aggregator can infer information about clients' training data from their model updates. To thwart such attacks, we augment FLGUARD with state-of-the-art secure computation techniques that securely evaluate the FLGUARD algorithm. We provide formal argumentation for the effectiveness of our FLGUARD and extensively evaluate it against known backdoor attacks on several datasets and applications (including image classification, word prediction, and IoT intrusion detection) demonstrating that FLGUARD can entirely remove backdoors with a negligible effect on accuracy. We also show that private FLGUARD achieves practical runtimes.**



***Index Terms*—federated machine learning, backdoor defenses, poisoning attacks, inference attacks**


## I. INTRODUCTION

*Federated learning* (FL) is an emerging collaborative machine learning trend with many applications such as next word prediction for mobile keyboards [1], medical imaging [2], and intrusion detection for IoT [3]. In FL, clients locally train model updates based on local training data (and a global model), and then provide these updates to a central aggregator who combines them to a new global model which is sent back to the clients for the next training iteration. FL offers efficiency and scalability as the training is distributed among many clients and executed in parallel.

In particular, FL improves privacy by enabling clients to keep their training data locally [4], which is not only relevant for compliance to privacy regulations such as the GDPR, but also in general when processing personal and sensitive data.

Despite its benefits, FL has been shown to be vulnerable to *backdoor* [5]–[7] and *inference* attacks [8]–[10]. In the former, the adversary stealthily manipulates the global model so that attacker-chosen inputs result in wrong predictions chosen by the adversary. The adversary can also inject multiple backdoors simultaneously. In the latter, the adversary aims at learning information about the clients' training local data by analyzing their model updates. Mitigating both attack types simultaneously is highly challenging: Backdoor defenses require access to the clients' model updates, whereas inference mitigation strategies prohibit this to avoid information leakage. Currently, there exists no solution that can adequately tackle both challenges at the same time.
Additionally, existing defenses against backdoor attacks are based on two main ideas: model clustering [11], [12] for identifying potentially poisoned model updates, as well as differential privacy-based techniques [13] clipping model weights and adding noise for diluting the impact of potentially poisoned model updates on the aggregated global model. The problem with the proposed clustering approaches is, however, that they fail to detect poisoned model updates in cases where multiple different backdoors are injected by the adversary simultaneously, or, when the adversary seeks to avoid detection, e.g., by scaling down the backdoor [5]. Differential privacy-based defenses, on the other hand, are effective in eliminating the impact of malicious model updates. However, they very often severely impact the benign performance of the aggregated model, as the clipping factors needed to ensure effective poison elimination also end up significantly modifying individual weights of benign model updates [5], [14].
Our solution, FLGuard overcomes these deficiencies by combining a novel approach for model clustering and an automated approach for selecting clipping and noising parameters. This results in negligible impact on the benign performance of the aggregated model, while ensuring effective elimination of malicious backdoors.

**Our Goals and Contributions.** In this paper, we propose the first FL system that is secure against state-of-the-art backdoor attacks while providing enhanced data privacy for the training data as follows:

1) *Security*: In contrast to several proposed defenses that are effective only against non-Byzantine adversaries in specific adversary models, our solution is generic in the sense that it can effectively mitigate backdoor attacks in a generic adversary model applicable to all known state-of-the-art backdoor attacks [5], [7], [11]. In particular, our defense considers also Byzantine adversaries. To the best of our knowledge, our solution is also the first that can address *multi-backdoor* injection attacks without significantly impacting the benign performance of the aggregated model. We provide formal argumentation and detail the design of our solution in Sect. III.

2) *Privacy*: By incorporating provably secure two-party computation techniques (STPC), FLGUARD prohibits access to local model updates, thus, impeding powerful inference attacks by a semi-honest aggregator. We detail the design of our efficient STPC protocols for each of FLGUARD's components in Sect. IV.

3) *Evaluation*: We extensively evaluate the efficiency and effectiveness of FLGUARD on various datasets such as word prediction, image classification and IoT intrusion detection (Sect. VI). We show that FLGUARD defeats all known backdoor attacks with a negligible effect on the accuracy of the aggregated model. We also show that the privacy-preserving variant of FLGUARD is practical, such that for training a neural network with 2.7 million parameters and 50 clients on CIFAR-10, it needs less than 13 minutes (Sect. VI-D).[1]

## II. BACKGROUND AND PROBLEM SETTING

### A. Federated Learning

In Federated learning (FL), $K$ clients and an aggregator $A$ collaboratively build a global model $G$ [4]. In training round $t \in [1, T]$ a subset of clients is chosen by the aggregator and each of these clients $i \in [1, K]$ locally trains a local model $W_i$ (with $p$ parameters/weights $w_i^1, \dots, w_i^p$) based on the previous global model $G_{t-1}$ using its local data $D_i$ and sends $W_i$ to $A$. Then, $A$ aggregates the models $W_i$ into the new global model $G_t$ by averaging the local models (weighted by the number of training samples used to train it): $G_t = \Sigma_{i=1}^{K} \frac{n_i \times W_i}{n}$, where $n_i = \|D_i\|, n = \Sigma_{i=1}^{K} n_i$. In practice, previous works employ equal weights ($n_i = n/K$) for the contributions of all clients [5], [7], [11]. We adopt this approach, i.e., we set $G_t = \Sigma_{i=1}^{K} \frac{W_i}{K}$.

### B. Poisoning Attacks on Federated Learning

When performing poisoning attacks, the adversary $\mathcal{A}^c$ manipulates the local models $W_i$ of $K' < \frac{K}{2}$ compromised clients to obtain poisoned models $W_i'$ which are then aggregated into the global model $G_t$ and affect its behavior. Poisoning attacks can be divided into *untargeted* and *targeted* attacks. In untargeted attacks, the adversary's goal is merely to impact (deteriorate) the benign performance of the aggregated model, while in targeted attacks (also called backdoor attacks), the adversary wants the poisoned model $G_t'$ to behave normally on all inputs except for specific attacker-chosen inputs $x \in I_{\mathcal{A}^c}$ (the so-called *trigger set*) for which attacker-chosen (incorrect) predictions should be output. Otherwise, the aggregator could merely reject such poisoned model updates that lead to a deterioration of the aggregated model's performance on benign data. In this paper, we therefore focus on targeted, i.e., backdoor attacks. To backdoor FL, the adversary can use data poisoning, e.g., [6], [7], [11] and/or model manipulation, e.g., [5], [15]:

– Data poisoning: $\mathcal{A}^c$ adds manipulated "poisoned" data to the training data used to train model $W_i'$.

– Model manipulation: $\mathcal{A}^c$ manipulates the training algorithms, its parameters, or directly manipulates (e.g., by scaling) the model $W_i'$.

When performing the attack, the adversary seeks to maximize attack impact while ensuring the distance (e.g., Euclidean distance) between poisoned models $W'$ and benign models $W$ remains below the detection threshold $\varepsilon$ of the aggregator: $\|W' - W\| < \varepsilon$. This is necessary to evade possible anomaly detection performed by the aggregator on individual clients' model updates.

**Adversary model (security).** Several earlier proposals for poisoning defenses make specific assumptions about adversarial behavior. They assume either that the adversary $\mathcal{A}^c$ behaves in a consistently malicious manner, contributing malicious model updates during all training rounds (e.g., [16]), or, limiting its attack to one particular trigger set, i.e., a single backdoor (e.g., [11]). Such assumptions are, however, not realistic since nothing prevents the adversary from displaying a Byzantine behavior: during each training round $t$, the adversary can either inject *no backdoors* (i.e., act as benign), one backdoor, or, multiple backdoors [5]. The adversary is also free to change its poisoning strategy at will. Due to this, earlier poisoning defenses are *not effective* against Byzantine adversaries and/or adversaries injecting multiple backdoors simultaneously. In this paper, we make therefore no specific assumptions about the behavior of the adversary. Further, we assume the adversary $\mathcal{A}^c$ to have full control over $K'$ ($K' < \frac{K}{2}$) clients and their training data, processes, and parameters [5], [7]. $\mathcal{A}^c$ also has full knowledge of the aggregator's operations, including potentially applied backdooring defenses. However, $\mathcal{A}^c$ has no control over any processes executed at the aggregator nor over the honest clients.

### C. Inference Attacks on Federated Learning

A number of attacks on FL have been proposed that aim at inferring information about the training data. White-box inference attacks extract information from the training gradients or parameters of a trained model or model update, while black-box inference attacks solely investigate the output of a trained model [17]. Furthermore, we can distinguish between *membership inference attacks* that try to determine the presence of a specific training sample in the training dataset [8], [9], [18], *property inference attacks* that try to extract properties of training samples [10], [18], *reconstruction attacks* that try to reconstruct samples from the training

[1]Upon acceptance of our paper, we will open source our implementation.

data [19], and *distribution estimation attacks* that try to assess the proportion of samples of a specific class in the data.

Previous works [17] found that white-box inference attacks by the aggregator $\mathcal{A}^s$ having access to every model update are significantly stronger than attacks on the aggregated global model. Furthermore, performing the attacks on the local models allows the adversary to link the gained information to a specific user, while the global model anonymizes the individual contributions. To summarize, an enhanced privacy protection for FL is needed that especially prohibits access to the local model updates.

**Adversary Model (privacy).** This adversary type, i.e., the honest-but-curious aggregator $\mathcal{A}^s$, has access to all local model updates $W_i$, and can thus perform model inference attacks on each local model $W_i$ to extract information about the corresponding participant's data $D_i$ used for training. $\mathcal{A}^s$ attempts to infer sensitive information about clients' data $D_i$ from their model updates $W_i$ [8]–[10], [18] by maximizing the information $\phi_i = \text{Infer}(W_i)$ that $\mathcal{A}^s$ gains about the data $D_i$ of client $i$ by inferring from its corresponding model $W_i$.

### D. Secure Two-Party Computation (STPC)

Secure Two-Party Computation (STPC) allows two parties to securely evaluate a function on their encrypted inputs. Thereby, the parties have only access to so-called secret-shares of the inputs that are completely random and therefore do not leak *any* information besides the final output. The real value can only be obtained if both shares are combined.

**Yao's Garbled Circuits (GC).** Yao introduced GCs [20] for STPC in 1986. The protocol is run between two parties called *garbler* and *evaluator*. The garbler generates the garbled circuit (GC) corresponding to the Boolean circuit to be evaluated securely by associating two random keys per wire that represent the bit values $\{0, 1\}$. The garbler then sends the GC together with the keys for his inputs to the evaluator. The evaluator obliviously obtains the keys for his inputs via Oblivious Transfer (OT) [21], and evaluates the circuit to obtain the output key. Finally, the evaluator maps the output key to the real output.

**Boolean/Arithmetic Sharing.** For every $\ell$-bit value $\mathsf{v}$, party $P_i$ for $i \in \{0, 1\}$ holds an additive sharing of the value denoted by $[\mathsf{v}]_i$ such that $\mathsf{v} = [\mathsf{v}]_0 + [\mathsf{v}]_1 \pmod{2^\ell}$. To securely evaluate a multiplication gate, the parties use Beaver's circuit randomization technique [22] where the additive sharing of a random arithmetic triple is generated in the setup phase [23]. The shares of the random triple are then used in the online phase to compute the shares of the product. In this line of work, the GMW protocol [24]–[26] takes a function represented as Boolean circuit and the values are secret-shared using XOR-based secret sharing (i.e., $\ell = 1$).

## III. Backdoor-resilient Federated Learning

To design our defense, we reviewed proposed defenses against backdoor attacks, as detailed in Sect. VII. They can be broadly divided into two categories. The first category of defenses aims at distinguishing between malicious and benign model updates [11], [12], [16], [27]. However, all of these approaches make very specific assumptions about the distribution of benign and malicious data, or, the characteristics of injected backdoors. As such assumptions do not hold in general, these defenses fail if any of these assumptions does not hold. In particular, none of the proposed approaches are able to reliably mitigate the simultaneous injection of multiple backdoors.

Another defense approach is inspired by differential privacy techniques and is based on 'diluting' the impact of poisoned models by clipping model weights and adding suitable noise [5] to individual clients' model updates. While these approaches can effectively mitigate backdoor impact on the global model, they have the significant drawback that they may radically reduce the benign main task accuracy, which is highly undesirable for any practical applications [5].

The above observations make it necessary to carefully design the poisoning defense framework in a way that combines the strengths of both approaches without suffering of their respective drawbacks.

### A. Design Rationale

We base our design on following theoretical considerations and observations:

**Observation III.1.** *Approaches utilizing model clustering with a fixed number $n$ of clusters for identifying malicious models are vulnerable to attacks with multiple backdoors $m \geq n$.*

Existing clustering-based defenses, e.g., [11], [12] aim to divide clients into $n = 2$ clusters: benign and malicious. However, by simultaneously injecting $m \geq n$ carefully crafted backdoors, the adversary can cause at least one or more of the $m$ backdoor models to be clustered together with benign models in the same cluster. Since the adversary can freely choose any $m \leq K'$, any clustering-based approach with a fixed number of clusters $n$ will likely not be effective in defending against all attacks.

Let us now denote the *discriminative ability* of a clustering approach, i.e., the smallest dissimilarity between two models $W_a$ and $W_b$ causing them to be classified into separate clusters with $\varepsilon$.

**Observation III.2.** *Adversary $\mathcal{A}^c$ can evade any clustering approach by ensuring that the distance between poisoned model updates $W'$ and benign models $W$ remains smaller than the discriminative ability $\varepsilon$ of the used clustering approach, i.e., if $\|W' - W\| < \varepsilon$. Models having a larger distance can be distinguished by the clustering approach and thus be filtered out.*

The discriminative ability $\varepsilon$ of any clustering approaches accepting more than one model per cluster will need to be tuned to a value that allows similar benign models $W$ to be clustered in the same cluster. Since models of individual benign clients are (in practical settings) trained based on non-identical training datasets, it follows that $\varepsilon > 0$.

This means that the adversary can evade anomaly detection by limiting the distance of its poisoned models to benign ones to be at most $\varepsilon$. For instance, Wang et al., [14] prove that no gradient-based algorithm can detect poisoned updates.

The goal of differential privacy-based defenses is to employ clipping of model weights and noising to individual model updates to limit the impact that potentially malicious weights can have on the resulting global model. This defense is parameterized by a clipping bound $\alpha$ and a noising level $\sigma$.

**Observation III.3.** *If the applied clipping bound $\alpha$ is too high, an adversary can boost its model $W'$ by scaling up its weights up to the clipping bound, thereby maximizing the impact on the aggregated global model $G_t$.*

**Observation III.4.** *If the applied clipping bound $\alpha$ is too low, also a large fraction of weights of benign model updates $W$ will undergo clipping, thereby leading to a deterioration of the accuracy of the resulting aggregated global model $G_t$ on benign data.*

It is therefore clear that the selection of the clipping bound $\alpha$ must be such that it minimizes potential impact of poisoned models while trying to avoid clipping benign models.

Given these considerations, we can postulate generic conditions for a secure poisoning defense. Let us denote with $\alpha_{opt}$ and $\sigma_{opt}$ an optimal clipping bound and noising level that will eliminate the impact of any poisoned model $W'$ having a maximum dissimilarity of $\varepsilon = \|W' - W\|$.

**Theorem III.5.** *Any combination of a clustering approach for model filtering and subsequent clipping and noising with parameters $\alpha_{opt}$ and $\sigma_{opt}$ will effectively mitigate backdoor attacks, if the discriminative ability of the clustering approach is smaller than or equal to $\varepsilon$.*

*Proof.* Under these conditions, poisoned models $W'$ having $\|W' - W\| > \varepsilon$ lie within the discriminative ability of the clustering approach and can thus be identified and eliminated. Poisoned models $W'$ for which $\|W' - W\| \leq \varepsilon$, lie outside the discriminative ability of the clustering approach and will thus not be filtered out. However, for such models, clipping and noising using parameters $\alpha_{opt}$ and $\sigma_{opt}$ will ensure that the contribution of such models is eliminated from the aggregated model $G_t$. $\square$

### B. Design Requirements

Based on the above observations we can define the requirements for an effective backdoor defense being able to mitigate multi-backdoor attacks in a generic setting: (1) A clustering-based approach must be able to handle simultaneous injection of multiple backdoors. To the best of our knowledge, none of the existing defenses achieve or even consider this. (2) Need to minimize $\varepsilon$ and optimize clipping and noising thresholds such that clipping and noising can work effectively.

Given that $\varepsilon_i = \|W' - W\|$, the difference between the poisoned global model $G'_t$ and the benign global model $G_t$ will be: $G'_t - G_t = \Sigma_{i=1}^{k} \frac{\varepsilon_i}{K}$. Therefore, in order to effectively mitigate the backdoor by using clipping and noising, we need to reduce $\Sigma_{i=1}^{k} \frac{\varepsilon_i}{K}$, by minimizing the number of poisoned models $k$ and $\varepsilon_i$ and then finding optimal clipping and noising thresholds $\alpha$ and $\sigma$. A naïve approach that merely combines existing techniques in an ad hoc "pipeline" does not work since straightforward application of existing approaches leads to ineffective results as we discuss in Sect. III-C and Sect.III-D, and evaluate in Sect. VI-B5. We need instead to consider three critical aspects: (1) Developing a new clustering approach capable of handling multiple simultaneous backdoors, (2) optimized parameter selection for clipping and noising, and (3) how to combine these approaches to achieve an effective defense.

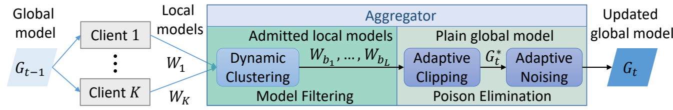



Fig. 1: Overview of FLGUARD in round $t$.

**Our solution.** For the first requirement, we introduce a dynamic clustering approach based on HDBSCAN. In contrast to existing approaches that aim at grouping all poisoned models into one group, our approach considers each poisoned model as an outlier. Therefore, our cluster approach can handle the simultaneous injection of multiple backdoors. For the second requirement, we first use our clustering approach to filter out model updates that have high attack impact, i.e., high $\varepsilon$ in order to reduce $k$, the number of poisoned model updates in the aggregation. We then introduce an automatic approach to specify a clipping bound and noise level based on benign data distribution. In particular, we use the median $L_2$-norm as a base for these parameters. Putting all together, we design a novel defense approach, FLGUARD, that has two layers: The first layer, called *Model Filtering* (Sect. III-C), uses dynamic clustering to identify and remove potentially poisoned model updates having high attack impact. The second layer, called *Poison Elimination* (Sect. III-D), leverages an adaptive threshold tuning scheme to clip model weights in combination with appropriate noising to smooth out and remove the backdoor impact of potentially surviving poisoned model updates.

### C. Filtering Poisoned Models

The *Model Filtering* layer utilizes a new dynamic clustering design that excludes models with high attack impact. It overcomes several limitations of existing defenses as (1) it can handle dynamic attack scenarios such as the simultaneous injection of multiple backdoors, and (2) it minimizes false positives.

Existing defenses [11], [12] cluster updates into two groups where the smaller group is always considered malicious and removed, leading to false positives and reduced accuracy when no attack is taking place. More importantly, $\mathcal{A}^c$ may also split compromised clients into groups injecting different backdoors. A fixed number of clusters bares the risk that poisoned and benign models end up in the same cluster, in particular, if models with different backdoors differ significantly. This is shown in Fig. 2 depicting different clusterings of model updates.

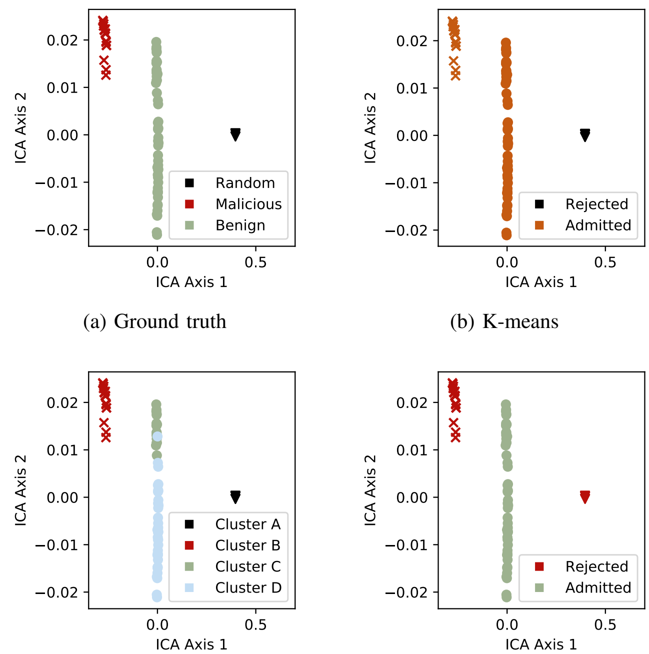
Fig. 2a shows the ground truth where $\mathcal{A}^c$ uses two groups of clients: 20 clients inject a backdoor and five provide random models to fool the deployed clustering-based defense. Fig. 2b shows how K-means (as used in [11]) fails to separate benign and poisoned models so that all poisoned ones end up in the same cluster with the benign models.

**Algorithm 1** FLGUARD

1: **Input:** $K$, $G_0$, $T$ ▷ $K$ is the number of clients, $G_0$ is the initial global model, $T$ is the number of training iterations
2: **Output:** $G_T$ ▷ $G_T$ is the updated global model after $T$ iterations
3: **for** each training iteration $t$ in $[1, T]$ **do**
4:     **for** each client $i$ in $[1, K]$ **do**
5:         $W_i \leftarrow$ CLIENTUPDATE$(G_{t-1})$ ▷ The aggregator sends $G_{t-1}$ to Client $i$ who trains $G_{t-1}$ using its data $D_i$ locally to achieve local modal $W_i$ and sends $W_i$ back to the aggregator.
6:     $(c_{11}, \ldots, c_{KK}) \leftarrow$ COSINEDISTANCE$(W_1, \ldots, W_K)$ ▷ $\forall i, j \in (1, \ldots, K)$, $c_{ij}$ is the Cosine distance between $W_i$ and $W_j$
7:     $(b_1, \ldots, b_L) \leftarrow$ CLUSTERING$(c_{11}, \ldots, c_{KK})$ ▷ $L$ is the number of admitted models, $b_l$ the indices of the admitted models
8:     $(e_1, \ldots, e_K) \leftarrow$ EUCLIDEANDISTISTANCES$(G_{t-1}, (W_1, \ldots, W_K))$ ▷ $e_i$ is the Euclidean distance between $G_{t-1}$ and $W_i$
9:     $S_t \leftarrow$ MEDIAN$(e_1, \ldots, e_K)$ ▷ $S_t$ is the adaptive clipping bound at round $t$
10:     **for** each client $l$ in $[1, L]$ **do**
11:         $W^*_{b_l} \leftarrow W_{b_l} * \text{MIN}(1, S_t/e_{b_l})$ ▷ $W^*_{b_l}$ is the admitted model after clipped by the adaptive clipping bound $S_t$
12:     $G^*_t \leftarrow \sum_{l=1}^{L} W^*_{b_l}/L$ ▷ Aggregating, $G^*_t$ is the plain global model before adding noise
13:     $\sigma \leftarrow \lambda * S_t$ ▷ Adaptive noising level
14:     $G_t \leftarrow G^*_t + N(0, \sigma)$ ▷ Adaptive noising

**Dynamic Clustering.** We overcome both challenges by calculating the pairwise Cosine distances measuring the angular differences between all model updates and applying the HDBSCAN clustering algorithm [29]. The Cosine distance is not affected by attacks that scale updates to boost their impact as this does not change the angle between the updates. While $\mathcal{A}^c$ can easily manipulate the $L_2$-norms of updates, reducing the Cosine distances decreases the attack impact [16]. HDBSCAN clusters the models based on their density and dynamically determines the required number of clusters. This can also be a single cluster, preventing false positives in the absence of attacks. Additionally, HDBSCAN labels models as noise if they do not fit into any cluster.

This allows FLGUARD to effectively handle multiple poisoned models with different backdoors by labeling them as noise to be excluded. We select the minimum cluster size to be at least 50% of the clients, i.e., $\frac{K}{2}+1$, such that it contains the majority of the updates (which we assume to be benign, cf. Sect. II-B). All remaining (potentially poisoned) models are marked as outliers. This behavior is depicted in Fig. 2d where the two benign clusters C and D from Fig. 2c are merged into one cluster while both malicious and random contributions are labeled as outliers. Hence, to the best of our knowledge, our clustering is the *first* FL backdoor defense for dynamic attacks where the number of injected backdoors varies. The clustering step is shown in Lines 6-7 of Alg. 1 where $L$ models $(W_{b_1}, \ldots, W_{b_L})$ are accepted.




(a) Ground truth (b) K-means

(c) HDBSCAN (d) FLGUARD

Fig. 2: Comparison of clustering quality for (a) ground truth, (b) using K-means with 2 clusters as in Auror [11], (c) naively applied HDBSCAN and (d) our approach as in FLGUARD. The models are visualized using Independent Component Analysis (ICA) [28].

**Personalization:** In some settings like next word suggestion [1], personalization might be an important aspect to power individual experiences, e.g., the style of writing. To personalize models, especially for benign clients whose model updates were rejected, each client, e.g., can improve re-train the final global model with its own data and use it immediately. For instance, Gboard, the Google Keyboard on Android, applies this approach to personalize text prediction based on what the user types on the keyboard [1].

### *D. Residual Poison Elimination by Smoothing*

To remove the impact of potentially remaining poisoned models that are not filtered by the *Model Filtering* layer (Sect. III-C), FLGUARD uses adaptive clipping and noising. In contrast to existing defenses that empirically specify a static clipping bound and noise level (and have been shown to be ineffective [5]), we automatically and adaptively tune these to effectively eliminate the backdoor's impact. Our design is also resilient to adversaries that dynamically adapt their attack.

Backdoor embedding makes poisoned models different from benign models. Clipping and noising can be combined to smooth model updates and remove these differences [30]. Clipping scales down the model weights to a clipping bound $S$. Noising refers to a technique that adds random noise to a model. While clipping and noising can remove backdoors, previous works [5] also show that they reduce the global model accuracy on the main task and potentially make it unusable. It is challenging to find an appropriate clipping bound $S$ and a noise level $\sigma$ that strikes a balance between the accuracy of the main task and effectiveness of the backdoor defense. Both need to be dynamically adapted to model updates in different training iterations and to different datasets (Sect. VI-B) as well as to dynamic adversaries constantly changing their attack strategy [5]. Note that this use of clipping and noising is different from differential privacy (DP; [30], [31]) protecting the confidentiality of clients' data from a curious aggregator and where clients truthfully train their models. In contrast, our scenario concerns malicious clients that intentionally try to backdoor FL. To overcome these challenges, we design our *Poison Elimination* layer for FLGUARD such that it automatically determines appropriate values for the clipping bound $S$ and the noise level $\sigma$.

**Adaptive Clipping.** Fig. 3 shows the variation of the average $L_2$-norms of model updates of benign clients in three different datasets over subsequent training rounds. This shows that the $L_2$-norms get smaller after each training iteration. To effectively remove backdoors while preserving benign updates unchanged, the clipping bound and noise level must dynamically adapt to this decrease in the $L_2$-norm. We design an adaptive selection of the clipping threshold $\alpha_{opt} = S_t$ for the $L_2$-norm for each training iteration $t$. The aggregator selects the median of the $L_2$-norms of the model updates $(W_1, \ldots, W_K)$ classified as benign in the clustering of our *Model Filtering* layer at iteration $t$. As we assume that the majority of clients is benign, this ensures that $S_t$ is determined based on a benign model even if some malicious updates were not detected during clustering. We formalize our clipping scheme as follows: $W^*_{b_l} = W_{b_l} * \text{MIN}(1, S_t/e_{b_l})$, where $S_t = \text{MEDIAN}(e_1, \ldots, e_L)$ in iteration $t$, see Lines 8-11 of Alg. 1 for details. By using the median, we ensure that the chosen clipping bound $S_t$ is always computed between a benign local model and the global model since we assume that more than 50% of clients are benign. We evaluate the effectiveness of our adaptive clipping approach in Sect. VI-B3.

**Adaptive noising.** We introduce a novel adaptive approach to calculate an appropriate level of noise based on the clipping bound $S_t$ in iteration $t$. We select the commonly used Gaussian distribution to generate noise that is added to the global model. Let $\sigma$ be the noise level and let $\lambda$ be a parameter indicating the product of $\sigma$ and the clipping bound $S_t$. Our adaptive noise addition is formalized as follows: $G_t = G^*_t + N(0, \sigma)$, where $\sigma = \lambda S_t$, for a clipping bound $S_t$ and a noise level factor $\lambda$, see Lines 13-14 of Alg. 1 for details.

In Sect. VI-B4, we empirically determine $\lambda = 0.001$ for image classification and word prediction, and $\lambda = 0.01$ for the IoT datasets.

## IV. Privacy-preserving Federated Learning

Inference attacks threaten the privacy of FL (cf. Sect. II-C). Especially when having access to the local model updates, the aggregator can infer sensitive information about the clients' training data with high accuracy [17]. Secure aggregation

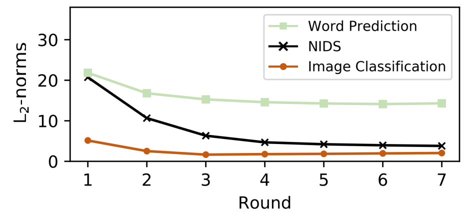



Fig. 3: $L_2$-norms depending on the number of training rounds for different datasets.

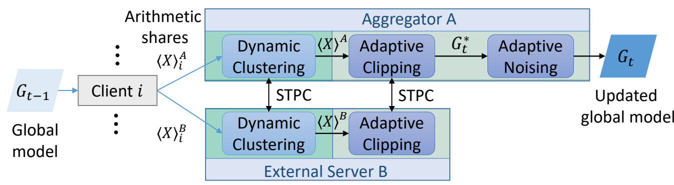



Fig. 4: Overview of private FLGUARD in round $t$ using Secure-Two-Party Computation (STPC).

protocols aggregate the updates in a secure manner such that the aggregator cannot use them for inference attacks. However, existing secure aggregation protocols [32]–[35] for FL solely focus on inference attacks and do not consider backdoor attacks and can typically not be combined with existing backdoor defenses (cf. Sect. VII).

### A. STPC for FL

Generally, there are two approaches to protect the privacy of clients' data: differential privacy (DP; [31]) and secure two-party computation (STPC; cf. Sect.II-D). DP is a statistical approach that can be efficiently implemented, but it can only offer high privacy protection at the cost of a significant loss in accuracy due to the noise added to the models [36], [37]. In contrast, STPC provides strong privacy guarantees and good efficiency but requires two non-colluding servers. Such servers can, for example, be operated by two competing companies that want to jointly provide a private FL service. To provide best efficiency and reasonable security, we chose semi-honest STPC for private FLGUARD. Alternatively, also more parties can be used in order to achieve better security at the cost of lower efficiency. These properties and assumptions are described and justified next.

**Semi-honest Security.** The semi-honest security model is standard in the security and privacy community [38]–[45] and can be justified by legal regulations such as the GDPR that mandate companies to properly protect users' data. Furthermore, service providers, e.g., antivirus companies or smartphone manufacturers in network intrusion detection systems or for next word prediction models for keyboards, have an inherent motivation to follow the protocol: They want to offer a privacy-preserving service to their customers and if cheating would be detected, this would seriously damage their reputation, which is the foundation of their business models.

**Instantiation.** To design the STPC protocols of FLGUARD, we use an optimized combination of three prominent STPC techniques: Yao's garbled circuits (GC, originally introduced by [20]) for the secure evaluation of non-linear operations in Boolean circuits in a constant number of rounds, as well as Boolean/Arithmetic sharing (originally introduced by [24]) for the secure evaluation of linear operations in Boolean/Arithmetic circuits with one round of interaction per layer of AND/Multiplication gates [24] (cf. Sect.II-D).

To co-design all components of FLGUARD as efficient STPC protocols, we generate a novel circuit for square root computation needed for determining cosine and $L_2$-norm distances using conventional logic synthesis tools. We carefully implement the circuit using Verilog HDL and compile it with the Synopsys Design Compiler [46] in a highly efficient way. We customize the flow of the commercial hardware logic synthesis tools to generate circuits optimized for GC including its state-of-the-art optimizations such as point-and-permute [47], free-XOR [48], FastGC [49], fixed-key AES [50], and half-gates [51]. For example, for the Free-XOR technique [48], which enables the evaluation of XOR gates without costly cryptographic encryption and thus makes GCs much more efficient, one has to minimize the number of non-XOR gates in the Boolean representation. We developed a technology library to guide the mapping of the logic to the circuit with no manufacturing rules defined similarly as in [52]–[54]. More concretely, to generate efficient Boolean circuits for FLGUARD, we constrained the mapping to free XOR gates and non-free AND gates. We enhanced the cost functions of the single gates: We set the delay and area of XOR gates to 0, the delay and area of the inverters to 0 (as they can be replaced with XOR gates with the constant input 1), and the delay and area of AND gates to a non-0 value. Note that the logic synthesis tool outputs a standard Boolean netlist containing cells that are included in the cell library. To use the netlist in a STPC framework [23], we performed post-synthesis. This circuit construction as well as the new circuit are also of independent interest. The new circuit can be used for other applications that need a privacy-preserving computation of square roots (e.g., any protocol that uses the Euclidean distance like privacy-preserving face recognition [55]). Moreover, the circuit construction chain is interesting for any other circuit that needs to be created and optimized for the GC protocol.

### B. *Private* FLGUARD

Fig. 5 shows the detailed processes of private FLGUARD. In 0, each client $i \in [1, K]$ determines its local model in a training round $t$. In 1, it splits the parameters of $W_i$ into two Arithmetic shares $\langle X\rangle_i^A$ and $\langle X\rangle_i^B$, such that $W_i = \langle X\rangle_i^A + \langle X\rangle_i^B$. The shares are sent to the aggregator A and the external server B over a secure channel.

Let $c_{ij}$ denote the Cosine distance (cf. Eq. 2 in Sect.C) between two models $W_i$ and $W_j$, where $i, j \in [1, K]$, and let $C = \{c_{11}, \ldots, c_{KK}\}$ be the set of all pairwise distances. In 2, $A$ and $B$ privately calculate the set $C$ and receive an arithmetic share of the set's elements as output, i.e., $A$ receives $\langle C\rangle^A = \{\langle c_{11}\rangle^A, \ldots, \langle c_{KK}\rangle^A\}$ and $B$ receives the respective $\langle C\rangle^B$. Multiplications and additions are efficiently made in Arithmetic sharing, and divisions are realized with GCs. A

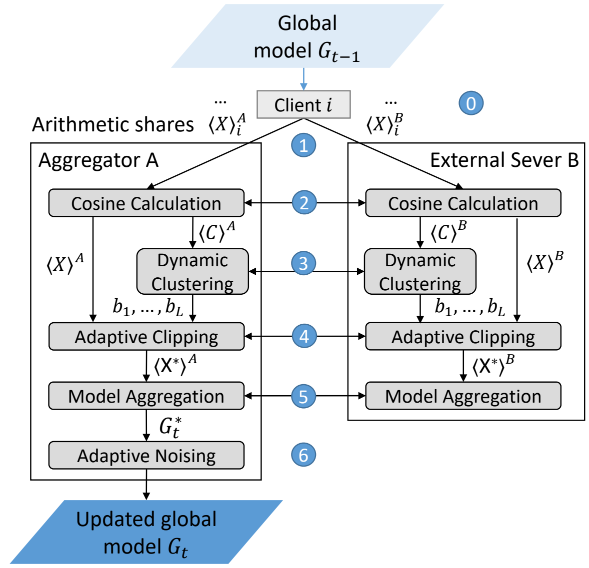



Fig. 5: Private FLGUARD processes in round $t$.

truncation is needed after each multiplication to preserve the size of the fractional part in fixed-point arithmetic. It can be efficiently realized with Boolean sharing, where the least significant bits are cut. This truncation method has on average a minor impact on the accuracy [38].

**Clustering.** In 3, clustering is applied to separate benign and malicious models based on similarities between the Cosine distances in $C$ (cf. Line 7 of Alg. 1). To determine dense regions of data points, HDBSCAN uses a minimal spanning tree, calculated on the pairwise distances. As the construction of the minimal spanning tree is expensive to realize with STPC [56], we use as approximation a privacy-preserving version of DBSCAN [57], a simplified version of HDBSCAN [29] that fixes the neighborhood notion to a maximum distance between two elements by using a parameter called $\epsilon$. The main difference between HDBSCAN and DBSCAN is that DBSCAN cannot handle clusters with varying densities very well, but as we create only a single cluster this is not problematic. We evaluate the accuracy of this approximation in Sect. VI-D. To determine an appropriate $\epsilon$-value, we conduct a binary search with several clusterings and varying $\epsilon$-values until one cluster contains exactly $\frac{K}{2}+1$ elements. This sacrifices some benign models that will wrongly be removed, but our evaluation in Sect. VI-D shows that private FLGUARD still successfully mitigates backdoors on all three datasets. Furthermore, this leaks only two bits of information to the servers, namely, if one cluster has the $\frac{K}{2}+1$ elements and if the boundary values for $\epsilon$ were changed. After determining the right $\epsilon$-value, a final clustering is executed and the resulting cluster indices are opened to A and B to enable them to determine the accepted models in 4. Moreover, A and B can also see who submitted a suspicious model but nothing about this client's training data. DBSCAN's second parameter, called $minPts$ and denoting the minimum cluster size, is set to $\frac{K}{2}$. The clustering outputs a list of clients with accepted models: $N = \{b_1, \ldots, b_L\}$, $L = \frac{K}{2} + 1$. For clustering, we purely rely on GC as it mainly works on binary values.

**Euclidean Distance, Clipping, and Model Aggregation.** Let $e_i$, $i \in \{1, \ldots, K\}$, denote the Euclidean distance between a model $W_i$ and the previous global model $G_{t-1}$ and let $E = \{e_1, \ldots, e_K\}$ indicate the set of these distances. In 4, A and B privately calculate $E$ such that A receives $\langle E \rangle^A = \{\langle e_1 \rangle^A, \ldots, \langle e_K \rangle^A\}$ and B receives the respective $\langle E \rangle^B$ as output. There, additions and multiplications are done in Arithmetic sharing, and square roots are calculated with GCs. Afterwards, each model $W_i$ is clipped based on its Euclidean distance $e_i$ to the previous global model $G_{t-1}$. To clip a model, the calculation of the median of Euclidean distances of the accepted models of the clients in $N$ is done with Boolean sharing and the division and the minimum determination are done with GCs. Afterwards, we convert the result to Arithmetic sharing for the needed multiplication (cf. Line 11 of Alg. 1). In 5, the clipped and accepted models are aggregated to the tentative model $G_t^*$. Arithmetic sharing is used for these summations. Then, in 6, $B$ sends its shares of $G_t^*$ to $A$ who reconstructs $G_t^*$ and divides it by $L$ before adding noise in plaintext. Using techniques from [58], we can also add noise in STPC to protect the global models at the expense of higher communication and computation. Finally, the new global model $G_t$ is sent back to the clients for the next training iteration.

## V. Experimental Setup

We implemented all experiments with the PyTorch framework and used the attack source code provided by Bagdasaryan et al. [5] and Xie et al. [7]. We reimplemented existing defenses to compare them with FLGuard. All experiments that evaluate FLGuard's effectiveness in defending backdoors were run on a server with 20 Intel Xeon CPU cores, 192 GB RAM, 4 NVIDIA GeForce GPUs (with 11 GB RAM each), and Ubuntu 18.04 LTS OS.

### A. Datasets and Learning Configurations

Following recent research on FL and poisoning attacks on FL, we evaluate our system in three typical application scenarios: word prediction [1], [4], [30], [59], image classification [2], [60], [61], and IoT [3], [6], [62]–[66]. Tab. I summarizes the used datasets and learning models.

TABLE I: Datasets used in our evaluations

| Application | WP | NIDS | IC | | |
|---|---|---|---|---|---|
| Datasets | Reddit | IoT-Traffic | CIFAR-10 | MNIST | Tiny-ImageNet |
| #Records | 20.6M | 65.6M | 60K | 70K | 120K |
| Model | LSTM | GRU | ResNet-18 Light | CNN | ResNet-18 |
| #params | ∼20M | ∼507K | ∼2.7M | ∼431k | ∼11M |

**Word Prediction.** We use the Reddit dataset of November 2017 [67] with the same parameters as [5] and [4], [30] for comparability. Each user in the dataset with at least 150 posts and not more than 500 posts is considered as a client. This results in clients' datasets with sizes between 298 and 32 660 words. The average client's dataset size is 4 111,6 words. We generated a dictionary based on the most frequent 50 000 words. The model consists of two LSTM layers and a linear output layer [4], [5]. It is trained for 5,000 iterations with 100 randomly selected clients in each iteration; each client trains for 250 epochs per iteration. The adversary uses 10 malicious clients to train backdoored models. To be comparable to the attack setting in [5], we evaluate FLGuard on five different trigger sentences corresponding to five chosen outputs (cf. Sect. F for the results).

**Image Classification.** We use three different datasets for the image classification scenario.

*CIFAR-10.* This dataset [68] is a standard benchmark dataset for image classification, in particular for FL [4] and backdoor attacks [5], [27], [69]. It consists of 60 000 images of 10 different classes. The adversary aims at changing the predicted label of one class of images to another class of images. [5] experiment with a backdoor where *cars in front of a striped background* are predicted to be *birds*, but we extend our evaluation to different backdoors, e.g., cats that are incorrectly labeled as airplanes (cf. Sect. G). We use a lightweight version of the ResNet18 model [70] with 4 convolutional layers with max-pooling and batch normalization [5].

*MNIST.* The MNIST dataset consists of 70 000 handwritten digits [71]. The learning task is to classify images to identify digits. The adversary poisons the model by mislabeling labels of digit images before using it for training [11]. We use a convolutional neural network (CNN) with

*Tiny-ImageNet.* Tiny-ImageNet[2] consists of 200 classes and each class has 500 training images, 50 validation images, and 50 test images. For Tiny-ImageNet, we used ResNet18 [70] as model.

**Network Intrusion Detection System (NIDS).** We test backdoor attacks on IoT anomaly-based intrusion detection systems that often represent critical security applications [3], [6], [72]–[76]. Here, the adversary aims at causing incorrect classification of anomalous traffic patterns, e.g., generated by IoT malware, as benign patterns. Based on the FL anomaly detection system DÏoT by [3], we use three datasets shared by [3] and [77] and one self-collected dataset from real-world home and office deployments located in Germany and Australia(cf. Sect. B). Following [3], we extracted device-type-specific datasets capturing the devices' communication behavior. Thereby, we prioritize device types that are present in several datasets and have sufficient data for evaluating them in a simulated FL setting where the data has to be split among the clients, i.e., *Security Gateways*. In total, we evaluate FLGuard on data from 50 devices of 24 device types. We simulate the FL setup by splitting each device type's dataset among several clients (from 20 to 200). Each client has a training dataset corresponding to three hours of traffic measurements containing samples of roughly 2 000-3 000 communication packets. We extensively evaluate FLGuard on all 13 backdoors corresponding to 13 Mirai's attacks (cf. Sect. D for details). However, by IoT-Traffic dataset we denote

[2]https://tiny-imagenet.herokuapp.com

a subset that contains data collected with the NetatmoWeather device type (a smart weather station). The model consists of 2 GRU layers and a fully connected output layer.

### B. Evaluations Metrics

We consider a set of metrics for evaluating the effectiveness of backdoor attack and defense techniques:

- **BA - Backdoor Accuracy** indicates the accuracy of the model in the backdoor task, i.e., it is the fraction of the trigger set for which the model provides the wrong outputs as chosen by the adversary. The adversary aims to maximize $BA$.
- **MA - Main Task Accuracy** indicates the accuracy of a model in its main (benign) task. It denotes the fraction of benign inputs for which the system provides correct predictions. The adversary aims at minimizing the effect on $MA$ to reduce the chance of being detected. The defense system should not negatively impact $MA$.
- **PDR - Poisoned Data Rate** refers to the fraction of poisoned data in the training dataset. Using a high $PDR$ can increase the $BA$ but is also likely to make poisoned models more distinguishable from benign models and thus easier to detect.
- **PMR - Poisoned Model Rate** is the fraction of poisoned models.
- **TPR - True Positive Rate** indicates how well the defense identifies poisoned models, i.e., the ratio of the number of models correctly classified as poisoned to the total number of models classified as poisoned.
- **TNR - True Negative Rate** indicates the ratio of the number of local models correctly classified as benign to the total number of models classified as benign. The higher the $TNR$, the less poisoned models are aggregated in the global model.

## VI. Experimental Results

### A. Preventing Backdoor Attacks

**Effectiveness of FLGuard.** We evaluate FLGuard against the state-of-the-art backdoor attacks called constrain-and-scale [5], DBA [5], and Edge-Case [14] using the same attack settings as in the original works with multiple datasets. The results are shown in Tab. II. FLGuard completely mitigates the constrain-and-scale attack ($BA = 0\%$) for all datasets. Moreover, our defense does not affect the main task performance of the system as the Main Task Accuracy ($MA$) reduces by less than $0.4\%$ in all experiments. The DBA attack as well as the Edge-Case attack [14] are also successfully mitigated ($BA = 3.2\%/4.0\%$).

We extend our evaluation to various backdoors on three datasets. For NIDS, we evaluate 13 different backdoors and 24 device types (cf. Sect. D and E), for word prediction 5 different word backdoors (cf. Sect. F), and for image classification 90 different image backdoors, which change the output of a whole class to another class (cf. Sect. G). In all cases, FLGuard successfully mitigates the attack while still preserving the $MA$.

TABLE II: Effectiveness of FLGuard against state-of-the-art attacks for the respective dataset, in terms of Backdoor Accuracy ($BA$) and Main Task Accuracy ($MA$).

| Attack | Dataset | No Defense | | FLGuard | |
|---|---|---|---|---|---|
| | | $BA$ | $MA$ | $BA$ | $MA$ |
| Constrain-and-scale [5] | Reddit | 100 | 22.6 | 0 | 22.3 |
| | CIFAR-10 | 81.9 | 89.8 | 0 | 91.9 |
| | IoT-Traffic | 100.0 | 100.0 | 0 | 99.8 |
| DBA [7] | CIFAR-10 | 93.8 | 57.4 | 3.2 | 76.2 |
| Edge-Case [14] | CIFAR-10 | 42.8 | 84.3 | 4.0 | 79.3 |
| Untargeted Poisoning [78] | CIFAR-10 | - | 46.72 | - | 91.31 |

TABLE III: Effectiveness of FLGuard in comparison to state-of-the-art defenses for the constrain-and-scale attack on three datasets, in terms of Backdoor Accuracy ($BA$) and Main Task Accuracy ($MA$).

| Defenses | Reddit | | CIFAR-10 | | IoT-Traffic | |
|---|---|---|---|---|---|---|
| | $BA$ | $MA$ | $BA$ | $MA$ | $BA$ | $MA$ |
| *Benign Setting* | - | 22.7 | - | 92.2 | - | 100.0 |
| *No defense* | 100.0 | 22.6 | 81.9 | 89.8 | 100.0 | 100.0 |
| Krum | 100.0 | 9.6 | 100.0 | 56.7 | 100.0 | 84.0 |
| FoolsGold | **0.0** | **22.5** | 100.0 | 52.3 | 100.0 | 99.2 |
| Auror | 100.0 | **22.5** | 100.0 | 26.1 | 100.0 | 96.6 |
| AFA | 100.0 | 22.4 | **0.0** | 91.7 | 100.0 | 87.4 |
| DP | 14.0 | 18.9 | **0.0** | 78.9 | 14.8 | 82.3 |
| FLGuard | **0.0** | 22.3 | **0.0** | **91.9** | **0.0** | **99.8** |

**Comparison to existing defenses.** We compare FLGuard to existing defenses: Krum [12], FoolsGold [16], Auror [11], Adaptive Federated Averaging (AFA; [27]), and a generalized differential privacy (DP) approach [5], [30]. Tab. III shows that FLGuard is effective for all 3 datasets, while previous works fail to mitigate backdoor attacks: $BA$ is mostly negligibly affected. Krum, FoolsGold, Auror, and AFA do not effectively remove poisoned models and $BA$ often remains at $100\%$. Additionally, the model's $MA$ is negatively impacted. These previously proposed defenses remove many benign updates (cf. Sect. VI-B) increasing the $PMR$ and rendering the attack more successful than without these defenses.

For example, Reddit's users likely provide different texts such that the distances between benign models are high while the distances between poisoned models are low as they are trained for the same backdoor. FoolsGold is only effective on the Reddit dataset ($TPR = 100\%$) because it works well on highly non-independent and identically distributed (non-IID) data (cf. Sect. VII). Similarly, AFA only mitigates backdooring on the CIFAR-10 dataset since the data are highly IID (each client is assigned a random set of images) such that the benign models share similar distances to the global model (cf. Sect. VII). The differential privacy-based defense is effective, but it significantly reduces $MA$. For example, it performs best on the CIFAR-10 dataset with $BA = 0$, but $MA$ decreases to $78.9\%$ while FLGuard increases $MA$ to $91.9\%$ which is close to the benign setting (no attacks), where $MA = 92.2\%$.

### B. Effectiveness of each of FLGuard's components

In this section, we separately evaluate the effectiveness of each of FLGuard's components.

*1) Resilience of our in-depth defense approach:* To evaluate the effectiveness of our combination of *Model Filtering* and *Poison Elimination*, we conduct experiments in which a

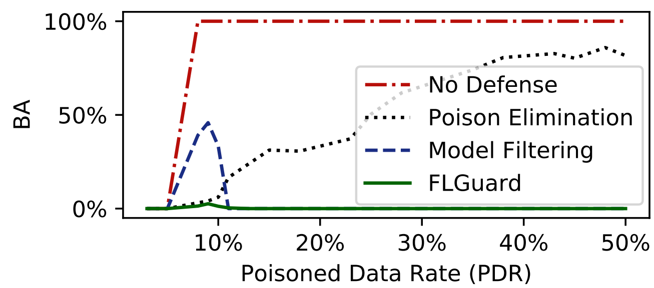



Fig. 6: Resilience of each defense layer in comparison to an effective combination in FLGUARD, measured by Backdoor Accuracy ($BA$).

sophisticated adversary can freely tune the attack parameter $PDR$ in order to find a setting that evades the filtering layer while still achieving a high $BA$. We show that the residual poisoned updates are eliminated by *Poison Elimination* in this case. We run experiments covering the full range of $PDR$ values to assess each defense component's effectiveness as well as the complete FLGUARD defense on the IoT-Traffic datasets. The Constrain-and-scale attack is used with the same settings as in Sect. VI-A.

Fig. 6 shows the $BA$ when using FLGUARD and its individual components depending on the $PDR$ values. As can be seen, *Model Filtering* can reliably identify poisoned models if $PDR$ is above $13\%$. Below this point, *Model Filtering* becomes ineffective as poisoned models become too indistinguishable from benign ones and cannot be reliably identified. Below this $PDR$ level, however, *Poison Elimination* can effectively remove the impact of poisoned models. Its performance only decreases when $PDR$ is increasing, and the impact of the backdoor functionality is harder to eliminate. However, our FLGUARD effectively combines both defense layers and remains successful for all $PDR$ levels as $BA$ consistently remains close to $0\%$.

*2) Effectiveness of the Clustering:* We show the results for the clustering in Tab. IV. As shown there, our clustering achieves $TNR = 100\%$ for the Reddit and IoT-Traffic datasets, i.e., FLGUARD only selects benign models in this attack setting. For the CIFAR-10 dataset, $TNR$ is not maximal (86.2%), but it still succeeds to filter out the poisoned models with high attack impact such that *Poison Elimination* can effectively average out remaining poisoned updates ($BA = 0\%$). Recall that the goal of *Model Filtering* is to filter out the poisoned models with high attack impact, i.e., not necessarily all poisoned models (cf. Sect. III).

**Impact of the Degree of non-Independent and Identically Distributed (non-IID) Data.** Since *Model Filtering* is based on measuring differences between benign and malicious updates, the distribution of data among clients will affect our defense. For CIFAR-10, we vary the degree of non-IID data, denoted by $Deg_{nIID}$, following previous work [78] by varying the fraction of images belonging to a specific class assigned to a specific group of clients. In particular, we divide the clients into 10 groups corresponding to the 10 classes of CIFAR-10. The clients of each group are assigned to a fixed fraction of $Deg_{nIID}$ of the images from its designated image class, while the rest of the images will be assigned to it at random. Consequently, the data distribution is random, i.e., completely IID if $Deg_{nIID} = 0\%$ (all images are randomly assigned) and completely non-IID if $Deg_{nIID} = 100\%$ (a client only gets images from its designated class). For the Reddit and IoT datasets, changing the degree of non-IID data is not meaningful since the data has a natural distribution as every client obtains data from different Reddit users or traffic chunks from different IoT devices. To summarize, our clustering approach provides almost identical results for different values of $Deg_{nIID}$ as $TNR$ and $TPR$ remain steady ($100.0\% \pm 0.00\%$ and $40.81\% \pm 0,00\%$), while $BA$ remains at $0\%$ and $MA$ is 91.9%($\pm$0.02%) for all experiments.

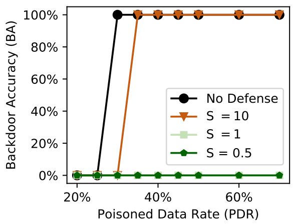
TABLE IV: Effectiveness of the clustering component, in terms of True Positive Rate ($TPR$) and True Negative Rate ($TNR$), of FLGUARD in comparison to existing defenses for the constrain-and-scale attack on three datasets. All values are in percentage and the best results of the defenses are marked in bold.

| Defenses | Reddit | | CIFAR-10 | | IoT-Traffic | |
|---|---|---|---|---|---|---|
| | TPR | TNR | TPR | TNR | TPR | TNR |
| Krum | 9.1 | 0.0 | 8.2 | 0.0 | 24.2 | 0.0 |
| FoolsGold | **100.0** | **100.0** | 0.0 | 90.0 | 32.7 | 84.4 |
| Auror | 0.0 | 90.0 | 0.0 | 90.0 | 0.0 | 70.2 |
| AFA | 0.0 | 88.9 | **100.0** | **100.0** | 4.5 | 69.2 |
| FLGUARD | 22.2 | **100.0** | 23.8 | 86.2 | **59.5** | **100.0** |

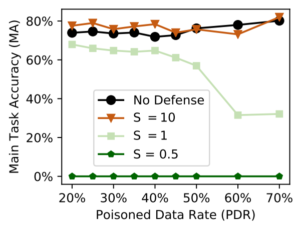
*3) Effectiveness of Clipping:* Fig. 7 demonstrates the effectiveness of FLGUARD's dynamic clipping where S is the $L_2$-norm median compared to a static clipping [5]. Fig. 7a and Fig. 7b show that a small static bound $S = 0.5$ is effective to mitigate the attack ($BA = 0\%$), but $MA$ drops to $0\%$ rendering the model inoperative. Moreover, a higher static bound like $S = 10$ is ineffective as $BA = 100\%$ if the Poisoned Data Rate ($PDR$) $\geq 35\%$. In contrast, FLGUARD's dynamic clipping threshold performs significantly better (cf. Fig. 7c and Fig. 7d). Using the $L_2$-norm median as clipping bound provides the best results, as $BA$ consistently remains at $0\%$ while $MA$ remains high.

*4) Effectiveness of Adding Noise:* Fig. 8 shows the impact of adding noise to the intermediate global models with respect to different noise level factors $\lambda$. As it can be seen, increasing $\lambda$ reduces the $BA$, but it also negatively impacts the performance of the model in the main task ($MA$). Therefore, the noise level must be dynamically tuned and combined with the other defense components to optimize the overall success of the defense.

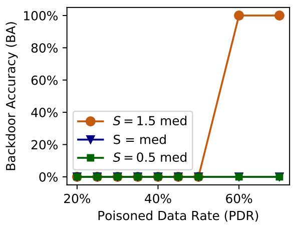
*5) Naïve combination:* Furthermore, we test a naïve combination of the defense layers by stacking clipping and adding noise (using a fixed clipping bound of 1.0 and a standard deviation of 0.01 as in [5]) on top of a filtering layer using K-means. However, this naïve approach still allows a $BA$ of 51.9% and a $MA$ of 60.24%, compared to a $BA$ of 0.0% and a $MA$ of 89.87% of FLGUARD in the same scenario. Based on our evaluations in Sect. VI-A, it becomes apparent



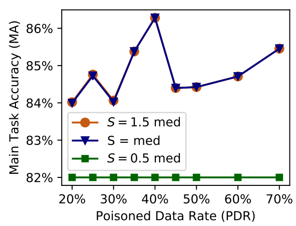

(a) BA for static bound




(b) MA for static bound

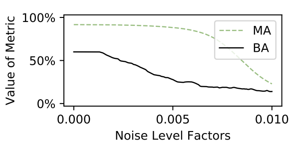



(c) BA for dynamic bound

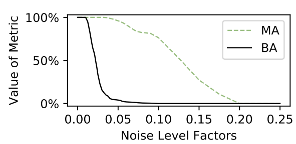



(d) MA for dynamic bound

Fig. 7: Effectiveness, in terms of Backdoor Accuracy ($BA$) and Main Task Accuracy ($MA$), of FLGUARD's dynamic clipping bound. $S$ is the clipping bound and $med$ the $L_2$-norm median.




(a) Image Classification




(b) Network Intrusion Detection

Fig. 8: Impact of different noise level factors on the Backdoor Accuracy ($BA$) and Main Task Accuracy ($MA$).

that FLGUARD's dynamic nature goes beyond previously proposed defenses that consist of static baseline ideas, which FLGUARD significantly optimizes, extends, and automates to offer a comprehensive dynamic and private defense against sophisticated backdoor attacks.

### C. Resilience to Adaptive Attacks

Given sufficient knowledge about FLGUARD, an adversary may seek to use adaptive attacks to bypass the defense layers. In this section, we analyze such attack scenarios and strategies including *changing the injection strategy*, *model alignment*, and *model obfuscation*.

**Changing the Injection Strategy.** The adversary may attempt to simultaneously inject several backdoors in order to execute different attacks on the system in parallel or to circumvent the clustering defense (cf. Sect. II-B). FLGUARD is also effective against such attacks (cf. Fig. 2 on p. 5). To further investigate the resilience of FLGUARD against such attacks, we conduct two experiments: (1) assigning different backdoors to malicious clients and (2) letting a malicious client inject several backdoors. We conduct these experiments with $K = 100$ clients of which $K' = 40$ are malicious on the IoT-Traffic dataset with each type of Mirai attack representing a backdoor. In the first experiment, we evaluate FLGUARD for $0, 1, 2,$ 4, and 8 backdoors meaning that the number of malicious clients for each backdoor is $0, 40, 20, 10,$ and 5. Our experimental results show that our approach is effective in mitigating the attacks as $BA = 0\% \pm 0.0\%$ in all cases, with $TPR = 95.2\% \pm 0.0\%$, and $TNR = 100.0\% \pm 0.0\%$. For the second experiment, 4 backdoors are injected by each of the 40 malicious clients. Also in this case, the results show that FLGUARD can completely mitigate the backdoors.

**Model Alignment.** Using the same attack parameter values, i.e., $PDR$ or $\alpha$ (cf. Sect. II-B), for all malicious clients can result in a gap between poisoned and benign models that can be separated by *Model Filtering*. Therefore, a sophisticated adversary can generate models that bridge the gap between them such that they are merged to the same cluster in our clustering. We evaluate this attack on the IoT-Traffic dataset for $K' = 80$ malicious clients and $K = 200$ clients in total. To remove the gap, each malicious client is assigned a random amount of malicious data, i.e., a random $PDR$ ranging from $5\%$ to $20\%$. Tab. V shows the effectiveness of FLGUARD against such attacks. Although FLGUARD cannot cluster the malicious clients well ($TPR = 5.68\%$), it still mitigates the attack successfully ($BA$ reduces from $100\%$ to $0\%$). This can be explained by the fact that when the adversary tunes malicious updates to be close to the benign ones, the attack's impact is reduced and consequently averaged out by *Poison Elimination*.

TABLE V: Resilience to model alignment attacks in terms of Backdoor Accuracy ($BA$), Main Task Accuracy ($MA$), True Positive Rate ($TPR$), True Negative Rate ($TNR$) in percent.

| | $BA$ | $MA$ | $TPR$ | $TNR$ |
|---|---|---|---|---|
| HDBSCAN | 100.0 | 91.98 | 0.0 | 33.04 |
| FLGUARD | **0.0** | **100.0** | **5.68** | **33.33** |

**Model Obfuscation.** The adversary can add noise to the poisoned models to make them difficult to detect. However, our evaluation of such an attack on the IoT-Traffic dataset shows that this strategy is not effective. We evaluate different noise levels to determine a suitable standard deviation for the noise. Thereby, we observe that a noise level of 0.034 causes the models' Cosine distances in clustering to change without significantly impacting $BA$. However, FLGUARD can still efficiently defend this attack: $BA$ remains at $0\%$ and $MA$ at $100\%$.

### D. Performance of Private FLGUARD

We evaluate the costs and scalability of FLGUARD when executed in a privacy-preserving manner by varying the number/size of the parameters that affect the three components realized with secure two-party computation (STPC) (cf. Sect. IV-B).

For our implementation, we use the ABY framework [23]. All STPC results are averaged over 10 experiments and run on two separate servers with Intel Core i9-7960X CPUs with

2.8 GHz and 128 GB RAM connected over a 10 Gbit/s LAN with 0.2 ms RTT.

**Runtime of Private FLGUARD.** Tab. VI shows the runtimes in seconds per training iteration of the Cosine distance, Euclidean distance + clipping + model aggregation, and clustering steps of Alg. 1 in standard (without STPC) and in private FLGUARD (with STPC). As can be seen, private FLGUARD causes a significant overhead on the runtime by a factor of up to three orders of magnitude compared to the standard (non-private) FLGUARD. However, even if we consider the largest model (Reddit) with $K = 100$ clients, we have a total server-side runtime of 22 081.65 seconds ($\approx$ 6 hours) for a training iteration with STPC. Such runtime overhead would be acceptable to maintain privacy, especially since mobile phones, which would be a typical type of clients in FL [4], are in any case not always available and connected so that there will be delays in synchronizing clients' model updates in FL. These delays can then also be used to run STPC. Furthermore, achieving provable privacy by using STPC may even motivate more clients to contribute to FL in the first place and provide more data.

**Communication of Private FLGUARD.** While in traditional FL each client sends its model to the server and later receives the aggregated model, in private FLGUARD, each client has to sent shares of its model to the two servers, and receives one aggregated model at the end. In addition, the communication in private FLGUARD is done using 64-bit fixed point numbers, while PyTorch uses 32-bit floating point numbers. Therefore, private FLGUARD increases the communication costs for each client by a factor of 3. In addition, also both aggregation servers need to communicate with each other. Tab. VII shows the communication costs of the servers in GB caused by using STPC for Cosine distance calculation, clustering, and Euclidean distance calculation/clipping/aggregating in each update iteration of FL. As the computation is done between two servers, we can assume a well-connected network with high throughput and low latency such that this overhead is acceptable.

**Approximating HDBSCAN by DBSCAN.** We measure the effect of approximating HDBSCAN by DBSCAN including the binary search for the neighborhood parameter $\epsilon$. The results are shown in Tab. VIII. As it can be seen, the results are very similar. For some applications, the approximation even performs slightly better than the standard FLGUARD. For example, for CIFAR-10, private FLGUARD correctly filters all poisoned models, while standard FLGUARD accepts a small number ($TNR = 86.2\%$), which is still sufficient to achieve $BA = 0.0\%$.

To conclude, private FLGUARD is the first privacy-preserving backdoor defense for FL with significant but manageable overhead and high effectiveness.

## VII. Related Work

**Backdoor Defenses.** Several backdoor defenses, such as Krum [12], FoolsGold [16], Auror [11], and AFA [27], aim at separating benign and malicious model updates. However, they only work under specific assumptions about the underlying data distributions, e.g., Auror and Krum assume that data of benign clients are independent and identically distributed (IID). In contrast, FoolsGold and AFA assume that benign data are non-IID. In addition, FoolsGold assumes that manipulated data is IID. As a result, these defenses are only effective under specific circumstances (cf. Sect. VI-A) and cannot handle the simultaneous injection of multiple backdoors (cf. Sect. III-C). In contrast, FLGUARD does not make any assumption about the data distribution and can defend against injection of multiple backdoors (cf. Sect. III-C).

Clipping and noising are known techniques to achieve differential privacy (DP) [31]. However, directly applying these techniques to defend against backdoor attacks is not effective because they significantly decrease the Main Task Accuracy (Sect. VI-A). FLGUARD tackles this by (i) identifying and filtering out potential poisoned models that have a high attack impact (cf. Sect. III-C), and (ii) eliminating the residual poison with an appropriate adaptive clipping bound and noise level, such that the Main Task Accuracy is retained (cf. Sect. III-D).

**Defenses against Inference Attacks in FL.** The first secure aggregation protocol by Bonawitz et al. [32] uses expensive additive masking and secret sharing to hide local updates. POSEIDON [33] encrypts the whole FL process resulting in a severe computational overhead. Other secure aggregation protocols also use either encryption [79] or secret sharing [34], [35], [80]. However, these protocols are all vulnerable to backdoor attacks as they prevent the aggregator from inspecting the model updates. BaFFLe [81] analyzes only the aggregated global models (to be compatible with secure aggregation) to detect backdoor injections based on feedback from clients, but it assumes that some clients hold data that triggers an atypical behavior caused by the backdoor injection. DP [30] limits the success of membership inference attacks. However, [18] has shown that this is only successful when thousands of clients are involved or for black-box attacks (cf. Sect.II-C). In private FLGUARD, local model updates are analyzed under STPC, thus, the aggregating servers cannot access the updates while thwarting backdooring.

## VIII. Summary

In this paper, we systematically analyzed backdoor attacks and introduced FLGUARD, the first FL defense that protects both privacy and security in a generic adversarial setting. Our extensive evaluation of various ML applications and datasets shows that FLGUARD can efficiently mitigate backdoor attacks without sacrificing the accuracy on the benign main task. Furthermore, we designed, implemented, and benchmarked efficient secure two-party computation protocols for FLGUARD to ensure privacy of clients' training data and to impede inference attacks on client updates.

## References


[1] B. McMahan and D. Ramage, "Federated learning: Collaborative Machine Learning without Centralized Training Data." Google AI, 2017.

[2] M. Sheller, A. Reina, B. Edwards, J. Martin, and S. Bakas, "Federated Learning for Medical Imaging," in *Intel AI*, 2018.

TABLE VI: Runtime in seconds of standard FLGUARD (S) in comparison to private FLGUARD using STPC (P). $K$ is the number of participating clients. Note that the model size has no effect on the clustering.

| K | Cosine Distance | | | | | | Euclidean Distance + Clipping + Model Aggregation | | | | | | Clustering | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reddit | | CIFAR-10 | | IoT-Traffic | | Reddit | | CIFAR-10 | | IoT-Traffic | | | |
| | (S) | (P) | (S) | (P) | (S) | (P) | (S) | (P) | (S) | (P) | (S) | (P) | (S) | (P) |
| 10 | 1.91 | 297.93 | 0.05 | 70.00 | 0.03 | 67.67 | 0.44 | 218.35 | 0.27 | 61.29 | 0.04 | 36.85 | 0.002 | 3.64 |
| 50 | 50.94 | 5 259.29 | 0.80 | 603.54 | 0.32 | 192.47 | 11.61 | 594.57 | 1.82 | 120.74 | 0.37 | 35.04 | 0.004 | 41.84 |
| 100 | 213.30 | 20 560.43 | 2.66 | 2 094.51 | 1.07 | 554.97 | 38.82 | 1 267.35 | 5.89 | 219.85 | 1.03 | 68.12 | 0.005 | 253.87 |

TABLE VII: Communication in GB of private FLGUARD's Cosine distances and of the Euclidean distances/clipping/model aggregation with different numbers of accepted models $\frac{K}{2}+1$ and different applications/model sizes. $K$ is the number of clients and clustering is independent of the model size.

| K | Cosine Distance | | | Euclidean Distance + Clipping + Model Aggregation | | | Clustering |
|---|---|---|---|---|---|---|---|
| | Reddit | CIFAR-10 | IoT-Traffic | Reddit | CIFAR-10 | IoT-Traffic | |
| 10 | 202 | 110 | 91 | 128 | 54 | 45 | 0.2 |
| 50 | 2 527 | 248 | 125 | 220 | 70 | 60 | 7.0 |
| 100 | 9 598 | 586 | 235 | 601 | 132 | 68 | 38 |

TABLE VIII: Effectiveness, in terms of Backdoor Accuracy ($BA$), Main Task Accuracy ($MA$), True Positive Rate ($TPR$), and True Negative Rate ($TNR$), of standard FLGUARD (S) in comparison to private FLGUARD using STPC (P) in percent.

| | Reddit | | CIFAR-10 | | IoT-Traffic | |
|---|---|---|---|---|---|---|
| | (S) | (P) | (S) | (P) | (S) | (P) |
| $BA$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| $MA$ | 22.3 | 22.2 | 91.9 | 91.7 | 99.8 | 99.7 |
| $TPR$ | 22.2 | 20.4 | 23.8 | 40.8 | 59.5 | 51.0 |
| $TNR$ | 100.0 | 100.0 | 86.2 | 100.0 | 100.0 | 100.0 |


[3] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A. Sadeghi, "DÏoT: A Federated Self-learning Anomaly Detection System for IoT," in *ICDCS*, 2019.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *AISTATS*, 2017.

[5] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," in *AISTATS*, 2020.

[6] T. D. Nguyen, P. Rieger, M. Miettinen, and A.-R. Sadeghi, "Poisoning Attacks on Federated Learning-Based IoT Intrusion Detection System," in *Workshop on Decentralized IoT Systems and Security*, 2020.

[7] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "DBA: Distributed Backdoor Attacks against Federated Learning," in *ICLR*, 2020.

[8] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "Knock Knock, Who's There? Membership Inference on Aggregate Location Data," in *NDSS*, 2018.

[9] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *IEEE S&P*, 2017.

[10] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property Inference Attacks on Fully Connected Neural Networks Using Permutation Invariant Representations," in *CCS*, 2018.

[11] S. Shen, S. Tople, and P. Saxena, "Auror: Defending Against Poisoning Attacks in Collaborative Deep Learning Systems," in *ACSAC*, 2016.

[12] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in *NIPS*, 2017.

[13] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963*, 2019.

[14] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *NeurIPS*, 2020.

[15] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," *IEEE S&P*, 2019.

[16] C. Fung, C. J. Yoon, and I. Beschastnikh, "The limitations of federated learning in sybil settings," in *RAID*, 2020, pp. 301–316, originally published as arxiv:1808.04866.

[17] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE S&P*, 2019.

[18] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning," in *IEEE S&P*, 2019.

[19] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *USENIX Security*, 2020.

[20] A. C.-C. Yao, "How to Generate and Exchange Secrets," in *FOCS*. IEEE, 1986.

[21] M. Naor and B. Pinkas, "Computationally Secure Oblivious Transfer," *Journal of Cryptology*, 2005.

[22] D. Beaver, "Efficient Multiparty Protocols Using Circuit Randomization," in *CRYPTO*, 1991.

[23] D. Demmler, T. Schneider, and M. Zohner, "ABY - A Framework for Efficient Mixed-Protocol Secure Two-Party Computation," in *NDSS*, 2015.

[24] O. Goldreich, S. Micali, and A. Wigderson, "How to Play any Mental Game," in *STOC*. ACM, 1987.

[25] A. Patra, T. Schneider, A. Suresh, and H. Yalame, "ABY2. 0: Improved mixed-protocol secure two-party computation," in *USENIX Security*, 2020.

[26] H. Yalame, H. Farzam, and S. Bayat-Sarmadi, "Secure two-party computation using an efficient garbled circuit by reducing data transfer," in *Applications and Techniques in Information Security*, 2017.

[27] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-Robust Federated Machine Learning through Adaptive Model Averaging," in *arXiv preprint:1909.05125*, 2019.

[28] C. Jutten and J. Herault, "Blind Separation of Sources: An Adaptive Algorithm Based on Neuromimetic Architecture," in *Signal Processing*, 1991.

[29] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2013.

[30] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning Differentially Private Language Models Without Losing Accuracy," in *ICLR*, 2018.

[31] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," in *Foundations and Trends in Theoretical Computer Science*, 2014.

[32] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *CCS*, 2017.

[33] S. Sav, A. Pyrgelis, J. R. Troncoso-Pastoriza, D. Froelicher, J.-P. Bossuat, J. S. Sousa, and J.-P. Hubaux, "POSEIDON: privacy-preserving federated neural network learning," in *NDSS*, 2021.

[34] J. H. Bell, K. A. Bonawitz, A. Gascón, T. Lepoint, and M. Raykova, "Secure single-server aggregation with (poly) logarithmic overhead," in *CCS*, 2020.

[35] J. So, B. Guler, and A. S. Avestimehr, "Turbo-aggregate: Break-

ing the quadratic aggregation barrier in secure federated learning," *arXiv:2002.04156*, 2020.
[36] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning," in *USENIX ATC*, 2020.
[37] Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving Deep Learning via Additively Homomorphic Encryption," in *TIFS*, 2017.
[38] P. Mohassel and Y. Zhang, "SecureML: A System for Scalable Privacy-Preserving Machine Learning," in *IEEE S&P*, 2017.
[39] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, "GAZELLE: A Low Latency Framework for Secure Neural Network Inference," in *USENIX Security*, 2018.
[40] P. Mishra, R. Lehmkuhl, A. Srinivasan, W. Zheng, and R. A. Popa, "DELPHI: A Cryptographic Inference Service for Neural Networks," in *USENIX Security*, 2020.
[41] J. Liu, M. Juuti, Y. Lu, and N. Asokan, "Oblivious Neural Network Predictions via MiniONN Transformations," in *CCS*, 2017.
[42] N. Agrawal, A. Shahin Shamsabadi, M. J. Kusner, and A. Gascón, "QUOTIENT: Two-Party Secure Neural Network Training and Prediction," in *CCS*, 2019.
[43] N. Kumar, M. Rathee, N. Chandran, D. Gupta, A. Rastogi, and R. Sharma, "CrypTFlow: Secure Tensorflow Inference," in *S&P*, 2020.
[44] M. S. Riazi, M. Samragh, H. Chen, K. Laine, K. Lauter, and F. Koushanfar, "XONN: XNOR-based Oblivious Deep Neural Network Inference," in *USENIX Security*, 2019.
[45] F. Boemer, R. Cammarota, D. Demmler, T. Schneider, and H. Yalame, "MP2ML: a mixed-protocol machine learning framework for private inference," in *ARES*, 2020.
[46] "Synopsys inc. design compiler," http://www.synopsys.com/Tools/Implementation/RTLSynthesis/DesignCompiler, 2010.
[47] D. Beaver, S. Micali, and P. Rogaway, "The Round Complexity of Secure Protocols," in *STOC*, 1990.
[48] V. Kolesnikov and T. Schneider, "Improved garbled circuit: Free XOR gates and applications," in *ICALP*, 2008.
[49] Y. Huang, D. Evans, J. Katz, and L. Malka, "Faster secure two-party computation using garbled circuits." in *USENIX Security*, 2011.
[50] M. Bellare, V. T. Hoang, S. Keelveedhi, and P. Rogaway, "Efficient garbling from a fixed-key blockcipher," in *sAndP*, 2013.
[51] S. Zahur, M. Rosulek, and D. Evans, "Two halves make a whole," in *Eurocrypt*, 2015.
[52] E. M. Songhori, S. U. Hussain, A.-R. Sadeghi, T. Schneider, and F. Koushanfar, "Tinygarble: Highly Compressed and Scalable Sequential Garbled Circuits," in *IEEE S&P*, 2015.
[53] D. Demmler, G. Dessouky, F. Koushanfar, A.-R. Sadeghi, T. Schneider, and S. Zeitouni, "Automated Synthesis of Optimized Circuits for Secure Computation," in *CCS*, 2015.
[54] M. Javadi, H. Yalame, and H. Mahdiani, "Small constant mean-error imprecise adder/multiplier for efficient VLSI implementation of MAC-based applications," in *TC*, 2020.
[55] M. Osadchy, B. Pinkas, A. Jarrous, and B. Moskovich, "SCiCI-A System for Secure Face Identification," in *IEEE S&P*, 2010.
[56] P. Laud, "Parallel Oblivious Array Access for Secure Multiparty Computation and Privacy-Preserving Minimum Spanning Trees," in *PETS*, 2015.
[57] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." in *KDD*, 1996.
[58] F. Eigner, A. Kate, M. Maffei, F. Pampaloni, and I. Pryvalov, "Differentially Private Data Aggregation with Optimal Utility," in *ACSAC*, 2014.
[59] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," in *ICLR*, 2018.
[60] M. Sheller, A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation," in *Brain Lesion Workshop*, 2018.
[61] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, "Project Adam: Building an Efficient and Scalable Deep Learning Training System," in *USENIX Operating Systems Design and Implementation*, 2014.
[62] J. Schneible and A. Lu, "Anomaly Detection on the Edge," in *IEEE Military Communications Conference*, 2017.
[63] J. Ren, H. Wang, T. Hou, S. Zheng, and C. Tang, "Federated Learning-Based Computation Offloading Optimization in Edge Computing-Supported Internet of Things," in *IEEE Access*, 2019.
[64] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Federated Learning for Ultra-Reliable Low-Latency V2V Communications," in *GLOBCOM*, 2018.
[65] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive Federated Learning in Resource Constrained Edge Computing Systems," in *JSAC*, 2019.
[66] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated Multi-Task Learning," in *NIPS*, 2017.
[67] "Reddit dataset," 2017, https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments.
[68] A. Krizhevsky and G. Hinton, "Learning Multiple Layers of Features from Tiny Images," Tech. Rep., 2009.
[69] M. Baruch, G. Baruch, and Y. Goldberg, "A Little Is Enough: Circumventing Defenses For Distributed Learning," in *NIPS*, 2019.
[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.
[71] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *IEEE*, 1998.
[72] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai Botnet," in *USENIX Security*, 2017.
[73] S. Herwig, K. Harvey, G. Hughey, R. Roberts, and D. Levin, "Measurement and Analysis of Hajime, a Peer-to-Peer IoT Botnet," in *NDSS*, 2019.
[74] R. Doshi, N. Apthorpe, and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," in *arXiv preprint:1804.04159*, 2018.
[75] S. Soltan, P. Mittal, and V. Poor, "BlackIoT: IoT Botnet of High Wattage Devices Can Disrupt the Power Grid," in *USENIX Security*, 2018.
[76] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and Other Botnets," in *IEEE Computer*, 2017.
[77] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, "Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics," in *TMC*, 2018.
[78] M. Fang, X. Cao, J. Jia, and N. Zhenqiang Gong, "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning," in *USENIX Security*, 2020.
[79] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, and Y. Zhou, "A Hybrid Approach to Privacy-preserving Federated Learning," in *AISec*, 2019.
[80] S. Kadhe, N. Rajaraman, O. O. Koyluoglu, and K. Ramchandran, "Fastsecagg: Scalable secure aggregation for privacy-preserving federated learning," *arXiv:2009.11248*, 2020.
[81] S. Andreina, G. A. Marson, H. Möllering, and G. Karame, "BaFFLe: backdoor detection via feedback-based federated learning," in *arXiv preprint:2011.02167*, 2020.
[82] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," in *ICML*. PMLR, 2018.

## APPENDIX

### A. Federated-Averaging Algorithm

The FedAvg aggregation rule is formalized in Alg. 2. Alg. 3 describes the client part of the training in FL.

**Algorithm 2** FedAvg (Aggregator-side execution)

**Input:** $K$, $G_0$, $T$ ▷ $K$ is the number of clients, $G_0$ is the initial global model, $T$ is the number of training iterations
**Output:** $G_T$ ▷ $G_T$ is the global model after $T$ iterations
**for** each training iteration $t$ in $[1, T]$ **do**
  **for** each client $i$ in $[1, K]$ **do**
    $W_i \leftarrow \text{CLIENTUPDATE}(G_{t-1})$ ▷ The Aggregator sends $G_{t-1}$ to Client $i$. The client trains $G_{t-1}$ using its data $D_i$ locally to achieve $W_i$ and sends $W_i$ back to the Aggregator.
  $G_t \leftarrow \sum_{i=1}^{K} n_i W_i / n$ ▷ Aggregating

**Algorithm 3** LocalTrain

```
▷ Once Client i receives G_{t-1}, it triggers LocalTrain(G_{t-1}, D_i) using its data D_i and sends W_i back to the Aggregator
function LocalTrain(G_{t-1}, D_i)
    W_i ← G_{t-1}
    for each batch b ⊂ D_i do
        W_i ← W_i − η∇ℓ(b, W_i)  ▷ ∇ℓ(b, W_k) denotes the gradient of the loss function ℓ for a training data batch b and η is the used learning rate
    return W_i
```

### B. IoT Datasets

Three datasets called DIoT-Benign, DIoT-Attack, and UNSW-Benign [3], [77] were kindly provided by the authors and consist of real-world traffic from four homes and two offices located in Germany and Australia. DIoT-Attack contains the traffic of 5 anomalously behaving IoT devices, infected by the Mirai malware [3]. Moreover, we collected a fourth IoT dataset containing communication data from 24 typical IoT devices (including IP cameras and power plugs) in three different smart home settings and an office setting. Table IX provides the details of the IoT datasets used in our experiments. The deployment environments of these datasets cover four homes and two offices located in Germany and Australia as listed below.

TABLE IX: Characteristics of IoT datasets

| Dataset | No. devices | Time (hours) | Size (MB) | Packets (millions) |
|---|---|---|---|---|
| FLGUARD-*Benign* | 28 | 7 603 | 1 153 | 6.4 |
| *DIoT-Benign* | 18 | 2 352 | 578 | 2.3 |
| *UNSW-Benign* | 27 | 7 457 | 11 759 | 23.9 |
| *DIoT-Attack* | 5 | 80 | 7 734 | 21.9 |

1) FLGUARD-*Benign*: Traffic that we captured from 28 IoT devices in three different smart home settings and an office setting. The smart home settings consists of two flats and one house in different cities, with 1 to 4 inhabitants. The office setting is a 20 $m^2$ office for two people. In each experiment, we deployed 28 IoT devices for more than one week and encouraged users to use the devices as part of their daily activities.
2) *DIoT-Benign*: Traffic that was captured from 18 IoT devices deployed in a real-word smart home [3].
3) *UNSW-Benign*: Traffic that was captured from 28 IoT devices in an office for 20 days [77].
4) *DIoT-Attack*: Traffic generated by 5 IoT devices infected by the Mirai malware [3].

Table X shows an overview of the 24 device types (78 devices in total) used during the evaluation of FLGUARD.

### C. Model Similarity Measures

Two measures are commonly used for evaluating the similarity between models: the $L_2$-norm (Euclidean distance) and the Cosine distance. A model $W = (w^1, w^2, \ldots, w^p)$ consists of $p$ model parameters $w^k, k \in [0, p]$. The similarity measures between two models $W_i$ and $W_j$, where $0 \le i, j \le K$ and $K$ is the number of clients, can therefore be defined as follows:

TABLE X: 24 device types used in the FLGUARD-*Benign*, *DIoT-Benign*, *UNSW-Benign*, and *DIoT-Attack* datasets.

| # | Device type | FLGUARD-Benign | DIoT-Benign | UNSW-Benign | DIoT-Attack |
|---|---|---|---|---|---|
| *1* | AmazonEcho | ○ | • | • | ○ |
| *2* | DLinkCam | • | • | ○ | • |
| *3* | DLinkType05 | • | • | ○ | ○ |
| *4* | EdimaxPlug | • | • | ○ | • |
| *5* | EdnetGateway | ○ | • | ○ | ○ |
| *6* | GoogleHome | ○ | • | ○ | ○ |
| *7* | HPPrinter | ○ | ○ | • | ○ |
| *8* | iHome | ○ | ○ | • | ○ |
| *9* | LightBulbsLiFXSmartBulb | ○ | ○ | • | ○ |
| *10* | Lightify2 | ○ | • | ○ | ○ |
| *11* | NestDropcam | ○ | ○ | • | ○ |
| *12* | NetatmoCam | • | ○ | • | ○ |
| *13* | NetatmoWeather | • | • | • | ○ |
| *14* | PIX-STARPhoto-frame | ○ | ○ | • | ○ |
| *15* | RingCam | • | ○ | ○ | ○ |
| *16* | SamsungSmartCam | ○ | ○ | • | ○ |
| *17* | Smarter | • | • | ○ | ○ |
| *18* | SmartThings | ○ | ○ | • | ○ |
| *19* | TesvorVacuum | • | ○ | ○ | ○ |
| *20* | TP-LinkDayNightCloudCamera | ○ | ○ | • | ○ |
| *21* | TPLinkPlug | • | ○ | • | ○ |
| *22* | TribySpeaker | ○ | ○ | • | ○ |
| *23* | WithingsAuraSmartSleepSensor | ○ | ○ | • | ○ |
| *24* | WithingsSmartBabyMonitor | ○ | ○ | • | ○ |

**Definition 1** ($L_2$-norm Distance)**.** *The $L_2$-norm distance $dl_{ij}$ between two models $W_i$ and $W_j$ with $p$ parameters, where $0 \le i, j \le K$, is the root of the squared parameter differences and is defined as:*

$$dl_{ij} = \|W_i - W_j\| = \sqrt{\sum_{k=1}^{p} (w_i^k - w_j^k)^2}. \quad (1)$$

**Definition 2** (Cosine Distance)**.** *The Cosine distance $dc_{ij}$ between two models $W_i$ and $W_j$ with $p$ parameters, where $0 \le i, j \le K$, measures the angular difference between the models' parameters and is defined as:*

$$\begin{aligned} dc_{ij} &= 1 - \frac{W_i W_j}{\|W_i\| \, \|W_j\|} \\ &= 1 - \frac{\sum_{k=1}^{p} w_i^k w_j^k}{\sqrt{\sum_{k=1}^{p} {w_i^k}^2} \sqrt{\sum_{k=1}^{p} {w_j^k}^2}}. \end{aligned} \quad (2)$$

### D. Effectiveness of FLGUARD for Different Mirai Attack Types

To evaluate the performance of FLGUARD against different backdoors (in this case, different *Mirai* attacks), we take all 13 attack types available in the attack dataset [3] and try to inject them as backdoors. The adversary controls 25 out of 100 clients and uses a *PDR* of 50%. For each backdoor, the adversary applies the Constrain-and-scale attack (cf. Sect. II-B) for 5 rounds, while FLGUARD is used as defense. Tab.

XI shows the results. It is visible that FLGUARD is able to mitigate all backdoor attacks completely while achieving a high $MA = 99.8\%$.

TABLE XI: Comparison of the Backdoor Accuracy ($BA$) when injecting different backdoors while using (1) Poison Elimination, (2) Clustering, and (3) FLGUARD as defense (Main Task Accuracy $MA = 99.8\%$ for all cases in FLGUARD).

| Backdoor | Baseline | (1) | (2) | (3) |
|---|---|---|---|---|
| Dos-Ack | 100.0% | 53.5% | 0.0% | 0.0% |
| Dos-Dns | 100.0% | 17.9% | 0.0% | 0.0% |
| Dos-Greeth | 100.0% | 19.3% | 0.0% | 0.0% |
| Dos-Greip | 100.0% | 59.8% | 0.0% | 0.0% |
| Dos-Http | 100.0% | 24.1% | 0.0% | 0.0% |
| Dos-Stomp | 100.0% | 95.0% | 100.0% | 0.0% |
| Dos-Syn | 100.0% | 13.5% | 0.0% | 0.0% |
| Dos-Udp | 100.0% | 40.0% | 0.0% | 0.0% |
| Dos-Udp (Plain) | 100.0% | 100.0% | 0.0% | 0.0% |
| Dos-Vse | 100.0% | 54.9% | 0.0% | 0.0% |
| Infection | 17.0% | 4.3% | 25.4% | 0.0% |
| Preinfection | 50.2% | 7.4% | 0.0% | 0.0% |
| Scan | 100.0% | 46.9% | 0.0% | 0.0% |
| Average | 89.8% | 41.3% | 9.6% | 0.0% |

### E. *Effectiveness of* FLGUARD *for Different Device Types*

Tab. XII and XIII show the effectiveness of FLGUARD and each individual defense technique. It is compared to the baseline that is not applying any defense measures. Analogous to the experiments in Tab. XI, the adversary controls 25% of the clients and uses a PDR of 50% for applying the Constrain-and-scale (cf. Sect. II-B) to inject a backdoor for the Mirai scanning attack. The attack was run for 3 training iterations. As it can be seen, FLGUARD is able to completely eliminate all backdoors ($BA = 0\%$), while preserving the accuracy of the model on the main task, i.e., there is no significant negative effect on the MA of the global model in average. Moreover, FLGUARD also clearly outperforms other defenses strategies that apply only single components of FLGUARD.

### F. *Performance of* FLGUARD *for Different NLP Backdoors*

To demonstrate FLGUARD's general applicability, we use it to defend backdoor attacks on a next word prediction task with multiple different backdoors as shown in Tab. XIV:
**(1): ”delicious”** after the sentence ”pasta from astoria tastes”
**(2): ”bing”** after the sentence ”search online using”
**(3): ”expensive”** after the sentence ”barbershop on the corner is”
**(4): ”nokia”** after the sentence ”adore my old”
**(5): ”rule”** after the sentence ”my headphones from bose”

### G. *Performance of* FLGUARD *for Different image backdoors*

To demonstrate FLGUARD's general applicability and evaluate its performance in wider attack scenarios than the very specific backdoor of [5] (who changed the output for cars in front of a striped background to birds) we also conducted 90 additional experiments for backdooring image classification.

TABLE XII: Backdoor Accuracy ($BA$) when applying (1) Clipping and adding noise, (2) Clustering, and (3) FLGUARD as defense.

| DeviceType | Baseline | (1) | (2) | (3) |
|---|---|---|---|---|
| AmazonEcho | 100.0% | 43.3% | 0.0% | 0.0% |
| DLinkCam | 100.0% | 47.6% | 0.0% | 0.0% |
| DLinkType05 | 100.0% | 44.2% | 0.0% | 0.0% |
| EdimaxPlug | 100.0% | 24.1% | 0.0% | 0.0% |
| EdnetGateway | 100.0% | 100.0% | 0.0% | 0.0% |
| GoogleHome | 100.0% | 87.1% | 0.0% | 0.0% |
| HPPrinter | 100.0% | 100.0% | 0.0% | 0.0% |
| iHome | 100.0% | 100.0% | 0.0% | 0.0% |
| LiFXSmartBulb | 100.0% | 92.2% | 0.0% | 0.0% |
| Lightify2 | 100.0% | 100.0% | 0.0% | 0.0% |
| NestDropcam | 100.0% | 62.4% | 0.0% | 0.0% |
| NetatmoCam | 100.0% | 60.0% | 100.0% | 0.0% |
| NetatmoWeather | 100.0% | 94.6% | 100.0% | 0.0% |
| PIX-STARPhoto | 100.0% | 100.0% | 0.0% | 0.0% |
| RingCam | 100.0% | 86.8% | 0.0% | 0.0% |
| SamsungSmartCam | 100.0% | 85.3% | 0.0% | 0.0% |
| Smarter | 100.0% | 100.0% | 0.0% | 0.0% |
| SmartThings | 100.0% | 100.0% | 0.0% | 0.0% |
| TesvorVacuum | 100.0% | 100.0% | 0.0% | 0.0% |
| TP-LinkCam | 100.0% | 100.0% | 0.0% | 0.0% |
| TPLinkPlug | 100.0% | 100.0% | 0.0% | 0.0% |
| TribySpeaker | 100.0% | 100.0% | 0.0% | 0.0% |
| WithingsSleepS | 100.0% | 100.0% | 0.0% | 0.0% |
| WithingsBabyM | 100.0% | 80.0% | 100.0% | 0.0% |
| Average | 100.0% | 83.7% | 12.5% | 0.0% |

In these experiments, we test on all possible pairs of instances and try to change the predictions of one class to each other possible class. Here, FLGUARD reduces the attack impact from $BA = 53.92 \pm 27.51$ to $BA = 2.52 \pm 5.83$ in average. However, note that even after applying FLGUARD the $BA$ is not zero as the model does not perform perfectly on all images even if it is not under attack. Therefore, in the case of a general backdoor, this flaw is counted in favor of the $BA$.

### H. *Evaluation of* FLGUARD *against DBA*

We evaluated FLGUARD in the same setup as used by Xie et al. [7] (but FLGUARD is integrated) for 3 different datasets (CIFAR-10, MNIST, and Tiny-ImageNet). In each training round, 10 (out of 100) randomly selected clients act maliciously. Following the setup of Xie *et al.*, we used a model that was trained only on benign clients and continued the training for some rounds in case of the CIFAR-10 and MNIST dataset with our FLGUARD being deployed, before launching the attack. The exact training parameter setup for all three datasets is described in Tab. XV.

Tab. XVI contains the results of the DBA when deploying FLGUARD compared to the baseline scenario where no defense is deployed. It can be seen that FLGUARD successfully mitigates the attack for all three datasets while preserving the $MA$. However, the $BA$ is not $0\%$ even before the attack because the model mislabels some images (as the $MA$ is not $100\%$) and this mislabeling is counted in favor for the $BA$ when the predicted label is equal to the target label by chance.

TABLE XIII: Main task accuracy (MA) when applying (1) Clipping and adding noise, (2) Clustering and (3) FLGUARD as defense.

| DeviceType | Baseline | (1) | (2) | (3) |
|---|---|---|---|---|
| AmazonEcho | 99.5% | 91.6% | 100.0% | 97.1% |
| DLinkCam | 99.7% | 98.5% | 97.5% | 89.9% |
| DLinkType05 | 84.7% | 76.9% | 98.7% | 94.2% |
| EdimaxPlug | 99.3% | 98.0% | 99.3% | 97.6% |
| EdnetGateway | 100.0% | 100.0% | 100.0% | 100.0% |
| GoogleHome | 100.0% | 94.7% | 100.0% | 99.9% |
| HPPrinter | 86.6% | 85.2% | 68.0% | 68.0% |
| iHome | 93.1% | 93.1% | 93.3% | 93.2% |
| LiFXSmartBulb | 94.3% | 96.6% | 93.5% | 93.4% |
| Lightify2 | 100.0% | 100.0% | 100.0% | 100.0% |
| NestDropcam | 100.0% | 100.0% | 100.0% | 100.0% |
| NetatmoCam | 99.2% | 98.7% | 99.3% | 97.5% |
| NetatmoWeather | 100.0% | 100.0% | 99.6% | 100.0% |
| PIX-STARPhoto | 100.0% | 100.0% | 100.0% | 100.0% |
| RingCam | 96.1% | 95.0% | 96.1% | 95.4% |
| SamsungSmartCam | 100.0% | 99.5% | 100.0% | 99.7% |
| Smarter | 93.3% | 93.3% | 100.0% | 100.0% |
| SmartThings | 100.0% | 100.0% | 100.0% | 100.0% |
| TesvorVacuum | 100.0% | 100.0% | 100.0% | 100.0% |
| TP-LinkCam | 67.2% | 67.2% | 67.1% | 67.0% |
| TPLinkPlug | 97.7% | 96.5% | 99.9% | 98.6% |
| TribySpeaker | 95.3% | 90.7% | 88.7% | 76.9% |
| WithingsSleepS | 100.0% | 100.0% | 100.0% | 100.0% |
| WithingsBabyM | 100.0% | 100.0% | 56.2% | 100.0% |
| Average | 96.1% | 94.8% | 94.0% | 94.5% |

TABLE XIV: Main Task Accuracy ($MA$), Backdoor Accuracy ($BA$), True Positive Rate ($TPR$), and True Negative Rate ($TNR$) of FLGUARD for different NLP backdoors (all values in percentage).

| | No Defense | | FLGUARD | | | |
|---|---|---|---|---|---|---|
| Backdoor | $BA$ | $MA$ | $BA$ | $MA$ | TPR | TNR |
| ”delicious” | 100.0 | 22.6 | 0.0 | 22.3 | 22.2 | 100.0 |
| ”bing” | 100.0 | 22.4 | 0.0 | 22.3 | 20.4 | 100.0 |
| ”expensive” | 100.0 | 22.2 | 0.0 | 22.3 | 20.4 | 100.0 |
| ”nokia” | 100.0 | 22.4 | 0.0 | 22.0 | 20.4 | 100.0 |
| ”rule” | 100.0 | 22.3 | 0.0 | 22.0 | 20.4 | 100.0 |
| Average | 100.0 | 22.4 | 0.0 | 22.2 | 20.8 | 100.0 |

## I. Impact of Number of Clients

Figure 9 shows the efficiency of FLGUARD in defending backdoors on the DLinkType05 device type from the IoT dataset with respect to different numbers of clients $(5, 10, \ldots, 100)$. As shown, the $TPR$ significantly varies if only a few clients are involved. The reason is that falsely rejecting only a single benign model has a high impact on the $TPR$. However, if more clients are involved, all metrics are stable. This shows that the effectiveness of FLGUARD is not affected by number of clients.

TABLE XV: Parameter setup for the evaluation of FLGUARD against the DBA.

| | CIFAR-10 | MNIST | Tiny-ImageNet |
|---|---|---|---|
| Number of Pretrained Rounds | 200 | 10 | 20 |
| Rounds without Attack | 2 | 1 | 0 |
| Local Epochs of Benign Clients | 2 | 1 | 2 |
| Local Epochs of Malicious Clients | 6 | 10 | 10 |
| Learning Rate of Benign Clients | $10^{-1}$ | $10^{-1}$ | $10^{-3}$ |
| Learning Rate of Malicious Clients | $5 * 10^{-2}$ | $5 * 10^{-2}$ | $10^{-3}$ |

TABLE XVI: Main Task Accuracy ($MA$) and Backdoor Accuracy ($BA$) of FLGUARD against the DBA (all values in percentage).

| | CIFAR-10 | | | | MNIST | | | | Tiny-ImageNet | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | No Defense | | FLGUARD | | No Defense | | FLGUARD | | No Defense | | FLGUARD | |
| | $BA$ | $MA$ | $BA$ | $MA$ | $BA$ | $MA$ | $BA$ | $MA$ | $BA$ | $MA$ | $BA$ | $MA$ |
| Pretrained Model | 2.2 | 75.9 | 2.2 | 75.9 | 0.5 | 97.2 | 0.5 | 97.2 | 0.1 | 56.5 | 0.1 | 56.5 |
| Before First Attack | 2.4 | 77.4 | 2.4 | 76.0 | 0.5 | 97.3 | 0.5 | 97.2 | 0.1 | 56.5 | 0.1 | 56.5 |
| After Attack | 93.8 | 57.4 | 3.2 | 76.2 | 99.3 | 87.9 | 0.5 | 97.3 | 97.0 | 16.3 | 0.1 | 56.4 |

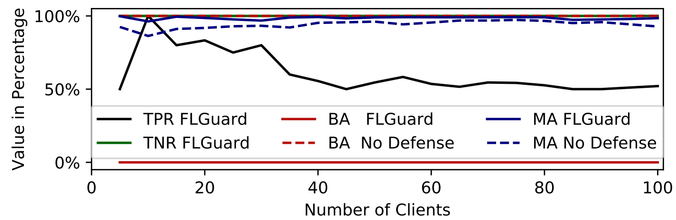



Fig. 9: Impact on the evaluation metrics of the total number of clients, using a fixed poisoned model rate $PMR$ =25%

## J. Impact of Number of Malicious Clients

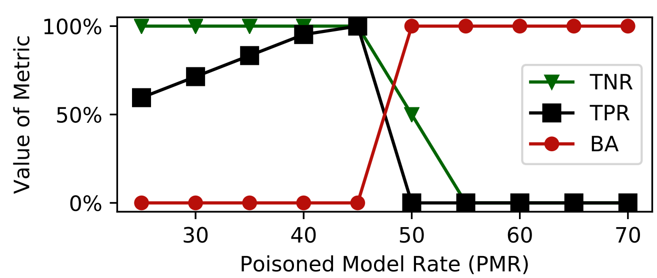



Fig. 10: Impact on the evaluation metrics of the poisoned model rate $PMR = \frac{K'}{K}$ which is the fraction of malicious clients $K'$ per total clients $K$.

We assume that more than half of all clients are benign (cf. Sect. II-B) and our clustering is only expected to be successful when $PMR = \frac{K'}{K} < 50\%$ (cf. Sect. III-C). We evaluate FLGUARD for different $PMR$ values. Fig. 10 shows how $BA$, $TPR$, and $TNR$ change in the NIDS application depending on $PMR$ values from $25\%$ to $75\%$. FLGUARD is only effective if $PMR < 50\%$ such that only benign clients are admitted to the model aggregation ($TNR = 100\%$) and thus $BA = 0\%$. However, if $PMR > 50\%$, FLGUARD fails to mitigate the attack because all malicious models will be included ($TPR = 0\%$).

Another attack type related to backdooring is *untargeted poisoning* resembling a denial of service (DoS) [12], [69], [78]. Unlike backdoor attacks that aim to incorporate specific backdoor functionalities, untargeted poisoning aims at rendering the model unusable. The adversary uses crafted local models with low Main Task Accuracy to damage the global model $G$. [78] propose such an attack bypassing state-of-the-art defenses. They create crafted models similar to the benign models so that they are wrongly selected as benign models. Although we do not focus on untargeted poisoning, our approach intuitively defends it since, in principle, this attack also trade-offs attack impact against stealthiness.

To evaluate the effectiveness of FLGUARD against untargeted poisoning, we test the sophisticated attack proposed by [78] on FLGUARD. The authors introduce three attacks against different aggregation rules: Krum [12], Trimmed Mean, and Median [82]. Among those three attacks, we consider the Krum-based attack because it: (1) is the focus of their work and stronger than the others, (2) can be transferred to unknown aggregation rules, and (3) has a formal convergence proof [12], [78]. Since [78]'s evaluation uses image datasets, we evaluate FLGUARD's resilience against it with CIFAR-10. Fig. 11 demonstrates FLGUARD's effectiveness against these untargeted poisoning attacks. It shows that although the attack significantly damages the model by reducing $MA$ from $92.16\%$ to $46.72\%$, FLGUARD can successfully defend against it and $MA$ remains at $91.31\%$.

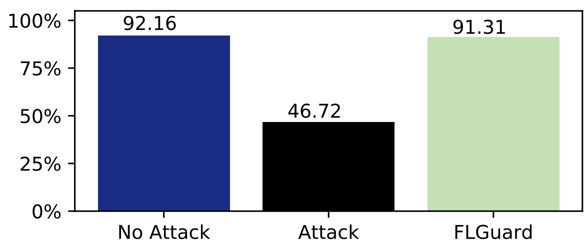



Fig. 11: Resilience of FLGUARD against an untargeted poisoning attack in terms of Main Task Accuracy ($MA$).

### K. Overhead of FLGUARD

We evaluated FLGUARD for 6 different device types from the IoT dataset (Amazon Echo, EdimaxPlug, DlinkType05, NetatmoCam, NetatmoWeather and RingCam). In this experiment, only benign clients participated and the model was randomly initialized. The highest observed overhead were 4 additional rounds. In average, $1.67$ additional training rounds were needed to achieve at least 99% of the $MA$ that was achieve without applying the defense.