META FEDERATED LEARNING

Omid Aramoon & Gang Qu
Department of Electrical and Computer Engineering
University of Maryland
{oaramoon,gangqu}@umd.edu

Pin-Yu Chen
IBM Research
{pin-yu.chen}@ibm.com

Yuan Tian
Department of Computer Science
University of Virginia
{yt2e}@virginia.edu

ABSTRACT

Due to its distributed methodology alongside its privacy-preserving features, Federated Learning (FL) is vulnerable to training time backdoor attacks. Contemporary defenses against backdoor attacks in FL require direct access to each individual client's update which is not feasible in recent FL settings where *Secure Aggregation* is deployed. In this study, we seek to answer the following question, *"Is it possible to defend against backdoor attacks when secure aggregation is in place?"*. To this end, we propose *Meta Federated Learning* (Meta-FL), a novel variant of FL which not only is compatible with secure aggregation protocol but also facilitates defense against backdoor attacks.

1 INTRODUCTION

Federated Learning (FL) is a distributed learning framework that enables millions of clients (e.g., mobile and edge devices) jointly train a deep learning model under the supervision of an orchestration server (McMahan et al. (2017); Smith et al. (2017); Zhao et al. (2018),§D.1). Taking advantage of the data distributed among the crowd of clients enables FL to train a highly accurate shared global model. In every round of FL, the central server randomly selects a cohort of participants to locally train the joint global model on their private data and submit an update to the server, which would be aggregated into the new global model. FL decouples model training from the need to access participants' data by collecting focused model updates that contain enough information to improve the global model without revealing too much about clients' private data (Kairouz et al., 2019).

While collecting model updates, instead of centralizing raw training data, significantly reduces privacy concerns for participating clients, it does not offer any formal privacy guarantees. Recent studies have shown that model updates can still leak sensitive information about the client's data (Melis et al., 2019; Nasr et al., 2018), which proves that preserving the privacy of clients is only a promise, and certainly not the reality of FL.

To address such privacy concerns, recent FL settings deploy Secure Aggregation (SecAgg) (Bonawitz et al. (2017),§D.2), a cryptographic protocol that enables the server to compute aggregate of updates and train the global model while keeping each individual update uninspectable at all time. Looking from the server's point of view, SecAgg can be a "double-edged sword." On the one hand, it can systematically mitigate privacy risks for participants, which would make FL more appealing to clients and eventually result in higher client turnout. On the other hand, it would facilitate training time adversarial attacks, such as backdoor attacks, by masking participants' contributions.

This paper answers the following question, *"Is it possible to defend against backdoor attacks when SecAgg is in place?"*, a question that has not been investigated by prior studies. To this end, we propose Meta Federated Learning (Meta-FL), a novel FL framework which not only preserves the privacy of participants but also facilitates defense against backdoor attacks. In our framework, we take full advantage of the abundance of participants by engaging more than one training cohort at each round to participate in model training. To preserve the privacy of participants, Meta-FL bootstraps the SecAgg protocol to aggregate updates from each training cohort. In Meta-FL, the

server is provided with a set of cohort aggregates, instead of individual model updates, which are further aggregated to generate the new global model. Please refer to Figure 3 in Appendix §A for an overview of model training in Meta-FL.

Meta-FL moves defense execution point from update level to aggregate level which facilitates mitigating backdoor attacks by offering the following advantages: (a) server can monitor cohort aggregates without violating the privacy of participants. Therefore, the adversary is forced to be mindful of their submissions and maintain stealth on the aggregate level as aggregates which are statistically different from others are likely to get flagged and discarded; (b) cohort aggregates exhibit less variation compared to individual client updates, which makes it easier for the server to detect anomalies, and (c) adversary faces competition from benign clients to hold control of the value of cohort aggregates which hinders them from executing intricate defense evasion techniques.

Our contributions are summarized as follows: **(i)** We propose meta federated learning, a novel FL framework that facilitates defense against backdoor attacks while protecting the privacy of participants, **(ii)** We show that moving the defense execution point from update level to aggregate level is effective in mitigating backdoor attacks without compromising privacy. **(iii)** We perform a systematic evaluation of contemporary defenses against backdoor attack in both baseline FL and Meta-FL. Results on two classification datasets of SVHN (Netzer et al., 2011) and GTSRB (Stallkamp et al., 2012) show that Meta-FL enhances the robustness of contemporary defense to backdoor attacks.

2 META FEDERATED LEARNING

In this section, we first discuss the challenges in mitigating backdoor attacks in FL. Then, we propose Meta Federated Learning (Meta-FL), and explain how it improves robustness to backdoor attacks while preserving the privacy of participating clients.

Challenges in defending against backdoor attack in FL are two-fold:
Challenge 1. Inspecting model updates is off limits with or without SecAgg. Recent studies have demonstrated that model updates can be used to partially reconstruct clients' training data (Yao et al., 2019; Li et al., 2019; Gu et al., 2019); therefore, any defensive approach which requires examination of submitted updates is a threat to the privacy of participants, and against privacy promises of FL. Moreover, inspecting model updates simply is not a valid option in systems augmented with SecAgg. Therefore, privacy promises of FL prohibit the server from auditing clients' submissions which gives the adversary the privilege to submit any arbitrary value without getting flagged as anomalous. We refer to this privilege as ***submission with no consequences***.
Challenge 2. Even without the restrictions mentioned above, defending against backdoor attacks would not be a trivial task. Model updates submitted by clients show high variations which makes it extremely difficult for the central server to identify whether an update works toward an adversarial goal. Sporadicity observed from model updates originates from the non-i.i.d distribution of the original dataset among participants, and the fact that each update is a product of stochastic gradient descent, a non-deterministic algorithm whose output is not merely a function of its input data.

Motivated to address the challenges above, we propose, Meta-FL, a novel federated setting which not only protects the privacy of clients but also aids the server in defense against backdoor attacks. Algorithm 1 summarizes different steps of model training in Meta-FL, which we discuss in details here. In each round t of training in Meta-FL, server randomly selects π cohorts $\{\zeta_1^t, \zeta_2^t, ..\zeta_\pi^t\}$, each containing c unique clients (Line 3). Training cohorts can be sampled in-order or independently. In the latter case, each cohort is sampled after another, and thus, no client will be a member of more than one cohort ($\zeta_i^t \cap \zeta_j^t = \emptyset$). In the recent case, there is no inter-dependency among cohort selection, and therefore, cohorts can have clients in common; this scenario is more suitable for cases where the number of participating clients is relatively small. Next, server broadcasts global model G^t to clients in each cohort (Line 5), each client i locally and independently trains the model G^t on their training data to obtain a new local model L_i^{t+1}, and compute their model update δ_i (Line 7). Then, the server establishes π separate instances of SecAgg protocol to concurrently compute the aggregate of updates submitted from clients of each cohort (Line 9). Finally, in the last stage of training, the server aggregates all the "cohort aggregates" using aggregation rule Γ, and updates the joint model with its learning rate η to obtain the next global model G^{t+1} (Line 11).

In our framework, plain model updates never leave the client's side. All participants are required to follow the SecAgg protocol and submit cryptography masked updates. SecAgg guarantees that the

server is able to aggregate the masked submissions to update the global model but can not obtain the value of each individual update. While each cohort aggregate may still leak information about collective training data of cohort members, the inferred information can not be associated with any individual client; therefore, the privacy of participants is preserved in Meta-FL.

Algorithm 1 Meta-FL framework

Initialize shared global model
for each round t in 1,2,3.. **do**
  Select π training cohorts $\{\zeta_1^t, \zeta_2^t, ..\zeta_\pi^t\}$.
  for cohort ζ_j^t in $\{\zeta_1^t, .., \zeta_\pi^t\}$ **in parallel do**
    Broadcast global model G^t.
    for client i in cohort ζ_j^t **in parallel do**
      $\delta_i^t \leftarrow ClientUpdate(i, G^t)$
    end for
    $\Delta_j^t \leftarrow SecAgg(\delta_1^t, \delta_2^t, .., \delta_c^t)$
  end for
  $G^{t+1} = G^t + \eta\, \Gamma(\Delta_1^t, \Delta_2^t, .., \Delta_\pi^t)$
end for

In Meta-FL, as the central server can only see the aggregate of training cohorts, defense mechanisms are obliged to carry out on aggregate level rather than update level. This property offers the server several advantages in mitigating backdoor attacks, which we will cover in the rest of this section. However, before we can proceed, we need to define several concepts that are key in understanding what follows. In the rest of this paper, we refer to a training cohort as adversarial if and only if there exists at least one malicious client among its members. Naturally, a cohort is referred to as benign if none of its members are malicious. Moreover, we refer to the aggregate of updates from a benign and an adversarial cohort as a benign and adversarial aggregate, respectively.

Moving defense execution point from update to aggregate level facilitates mitigating backdoor attacks by offering following advantages:

Advantage 1. Server is allowed to inspect and monitor cohort aggregates. This property forces the adversary to maintain stealth on the aggregate level as adversarial aggregates which are statistically different from other benign aggregates are likely to get detected and discarded by the server. Therefore, Meta-FL revokes the privilege of submission with no consequences.

Advantage 2. Cohort aggregates are less sporadic compared to individual client updates which aids the server in detecting anomalies. This advantage takes on the *challenge 2* discussed above. By drawing an analogy to *simple random sampling* in statistics (Rice, 2006), we demonstrate that cohort aggregates show less variation across each coordinate compared to individual updates.

For ease of analysis, we assume that training cohorts are sampled independently meaning that there is no inter-dependency among client selection in each cohort. In this case, at any round t, updates submitted by any cohort of c clients are essentially a random sample of size c collected without replacement from the population of model updates. Assuming that updates are averaged, cohort aggregates are in fact *sample means* of model update population. As the composition of cohorts is a random process, cohort aggregates are thus random variables whose distribution is determined by that of model updates as shown below (for proof refer to (Rice, 2006))

$$Var(\Delta_j) = \frac{\sigma_j^2}{c}\left(\frac{P-c}{P-1}\right), \quad \mathbb{E}[\Delta_j] = \mathbb{E}[\mu_j] \tag{1}$$

Here, σ_j^2 and μ_j denote variance and mean of population of model updates across the j_{th} coordinate, respectively, and Δ_j indicates the j_{th} coordinate of a cohort aggregate Δ. Assuming that each cohort contains more than one client ($1 < c$), it'd be trivial to show that $\frac{P-c}{c(P-1)} < 1$. Therefore, we can prove that the variance of aggregates across any coordinate j is upper bounded by the variance of model updates across that coordinate. Lower variation from cohort aggregates makes it easier for outlier detection-based defenses to infer patterns of benign observations, and effectively detect out-of-distribution malicious instances.

Advantage 3. As adversarial updates are aggregated with other updates, malicious clients face competition from benign clients to control the value of cohort aggregate. This property makes it harder for the adversary to meticulously arrange values of adversarial aggregates to evade defenses.

3 EVALUATION

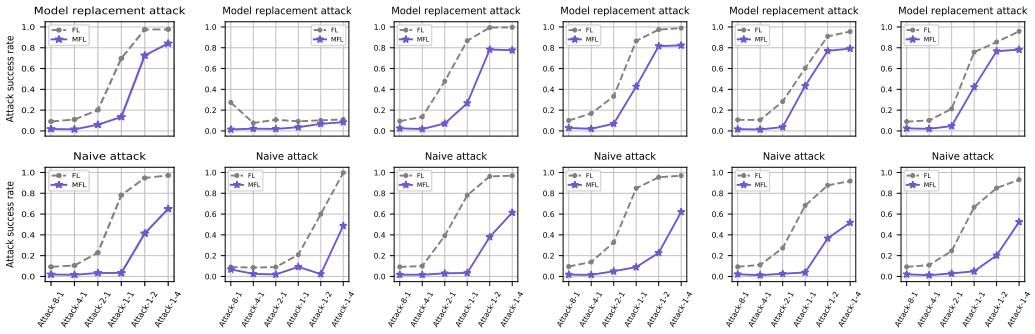
Experiment Setup. We study Meta-FL on two classification datasets namely SVHN (Netzer et al., 2011) and GTSRB (Stallkamp et al., 2012) with non-i.i.d. data distributions. For more details on

dataset description and their corresponding benchmark DNNs, and parameter setups, please refer to §B. Similar to the analysis in (Sun et al., 2019), we experiment with fixed frequency attack scenarios. In the baseline FL, **Attack-f-k** describes a scenario where k attackers appear at every f round of training. For the case of Meta-FL setting, **Attack-f-k** describes the case where at every f round of training, k training cohorts contain an adversarial client. In our threat model (§E), the adversary seeks to mount pixel pattern backdoor attack by adopting either the "Naive" or the "Model Replacement" techniques (Bagdasaryan et al., 2020). For more details on adversary's capabilities and objective refer to §E.

Meta-FL vs Baseline FL. In this section, we compare the capabilities of contemporary defenses against backdoor attacks in both baseline and Meta-FL. Our empirical evaluation shows that all defenses benefit from the advantages discussed in §2 and offer better robustness in Meta-FL.

We consider contemporary defenses such as Krum (Blanchard et al., 2017), Coordinate-Wise Median (CWM) and Trimmed Mean (TM) (Yin et al., 2018), Norm Bounding (NB) and Differential Privacy (DP) (Sun et al., 2019), and RFA (Pillutla et al., 2019). For description of these techniques refer to §C. The comparison is performed between a Meta-FL framework with 15 cohorts of 5 clients and a baseline FL with a single cohort of 15 clients. For this experiment, we set the number of training cohorts in Meta-FL equal to the number of selected clients in baseline FL to ensure that server sees the same number of "aggregands" (15 client updates in baseline FL and 15 cohort aggregates in Meta-FL) across both cases. Moreover, with the way our attack scenarios are designed, the same number of aggregands are adversarial across both frameworks, which assures a fair comparison.

Figures 1 and 2 report performance of contemporary defenses against backdoor attacks on GTSRB and SVHN benchmarks, respectively. As shown, ***Meta-FL puts all defense at an advantage in mitigating against backdoor attacks.*** Attack success rate of both the naive and model replacement approach in Meta-FL (solid lines) is lower than in baseline FL (dashed lines) when the same defense is in place across both frameworks. Therefore, our evaluations show that existing defenses are more robust to backdoor attacks in Meta-FL compared to baseline FL. Please refer to §F, for a more detailed analysis of experimental results.



(a) CWM (b) Krum (c) TM (d) RFA (e) NB (f) DP

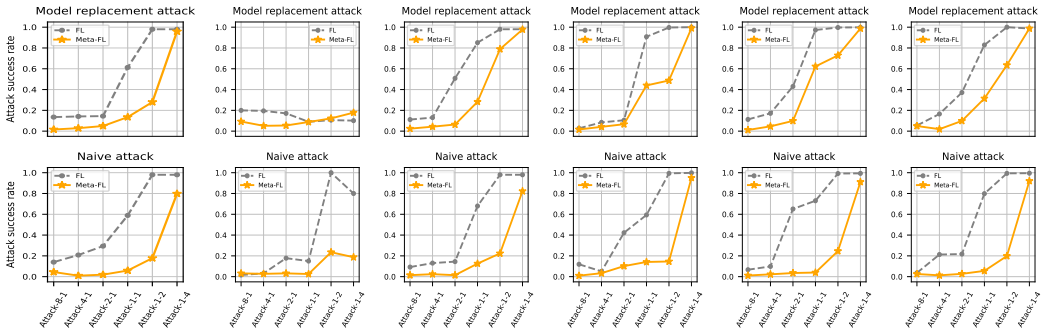
Figure 1: Performance of defenses against backdoor attacks in SVHN model training.

4 Conclusion

In this paper, we propose Meta-FL, a new FL framework which not only protects the privacy of participants through the SecAgg protocol but also facilitates defense against backdoor attacks. Our empirical evaluations demonstrate that contemporary defense tends to be more effective against backdoor attacks in Meta-FL compared to baseline FL.

References

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948. PMLR, 2020.



(a) CWM (b) Krum (c) TM (d) RFA (e) NB (f) GTSRB-DP

Figure 2: Performance of defenses against backdoor attacks in GTSRB model training.

Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pp. 634–643. PMLR, 2019.

Peva Blanchard, Rachid Guerraoui, Julien Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pp. 119–129, 2017.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *arXiv preprint arXiv:1909.02742*, 2019.

Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. *arXiv preprint arXiv:1903.09860*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706. IEEE, 2019.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. *arXiv preprint arXiv:1812.00910*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.

John A. Rice. *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third edition, 2006.

Claude E Shannon. Communication theory of secrecy systems. *The Bell system technical journal*, 28(4):656–715, 1949.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 4424–4434, 2017.

J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, (0):–, 2012. ISSN 0893-6080. doi: 10.1016/j.neunet.2012.02.016. URL http://www.sciencedirect.com/science/article/pii/S0893608012000457.

Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgyS0VFvr.

Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2041–2055, 2019.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*, 2018.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

A Overview of Meta-FL

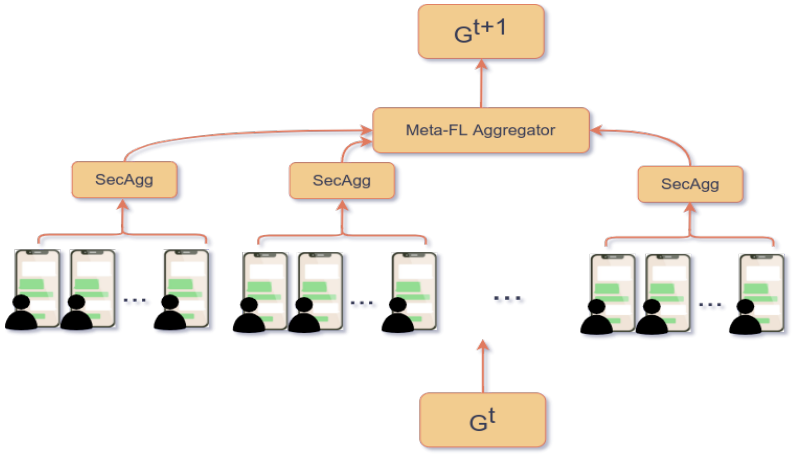


Figure 3: Overview of model training in meta federated learning.

B Datasets and Their Benchmark Neural Networks

GTSRB is s traffic sign dataset with 39,209 training and 12,630 test samples, where each sample is labeled with one of the 43 classes, and **SVHN** is a dataset of more than 100k images of digits cropped out of images of houses and street numbers. Table 1 reports the topology and hyperparameters of benchmarks used for GTSRB and SVHN datasets.

We use a Dirichlet distribution with parameter $\alpha = 0.9$ to partition GTSRB and SVHN datasets into disjoint non-i.i.d shards and then distribute them among 150 and 300 clients, respectively. Following a similar setup to prior arts, each participating client trains their local model using SGD for 5 epochs with a batch size of 64 and a learning rate of 0.1. Both Meta-FL and baseline FL resume the training process until a certain number of training rounds are completed. Throughout our experiments, GTSRB and SVHN models are trained for 75 and 50 rounds, respectively.

For all experiments, pixel pattern backdoor attacks are performed in which the adversary aims to influence the model to misclassify inputs from a base label as a target label upon the presence of an attacker chosen pattern (trigger). We set the adversarial trigger as a white square located at the top left corner of the image which roughly covers 9% of the entire image. The objective of backdoor attacks in GTSRB and SVHN datasets is to mispredict images of "Speed limit 80 miles per hour" as "Speed limit 50 miles per hour" and images of "digit 6" as "digit 1", upon the presence of the white box trigger.

Table 1: Model architecture for SVHN and GTSRB datasets.

GTSRB		SVHN	
Layer Type	Filter/Unit	Layer Type	Filter/Unit
Conv + ReLU	$3 \times 3 \times 32$	Conv + ReLU	$3 \times 3 \times 32$
Conv + ReLU	$3 \times 3 \times 32$	Conv + ReLU	$3 \times 3 \times 32$
Conv + ReLU	$3 \times 3 \times 64$	Conv + ReLU	$3 \times 3 \times 64$
Conv + ReLU	$3 \times 3 \times 64$	Conv + ReLU	$3 \times 3 \times 64$
Conv + ReLU	$3 \times 3 \times 128$	Conv + ReLU	$3 \times 3 \times 128$
Conv + ReLU	$3 \times 3 \times 128$	Conv + ReLU	$3 \times 3 \times 128$
FC + ReLU	43	FC + ReLU	512
Softmax	43	FC + ReLU	10
		Softmax	10

C ROBUST AGGREGATION RULES AND DEFENSES

Numerous studies have proposed robust aggregation rules (Blanchard et al., 2017; Yin et al., 2018; Pillutla et al., 2019) to ensure convergence of distributed learning algorithms in the presence of adversarial actors. The majority of studies in this line of work assume a byzantine threat model in which the adversary can cause local learning procedures to submit any arbitrary update to ensure convergence of learning algorithms to an ineffective model. In addition to robust aggregation rules, several works have proposed novel defenses (Fung et al., 2018; Sun et al., 2019) against backdoor and poisoning attacks in FL. In what follows, we review several of the techniques which we experiment in §3.

Krum. The Krum algorithm, proposed by (Blanchard et al., 2017), is a robust aggregation rule which can tolerate f byzantine attackers out of n participants selected at any training round. Krum has theoretical guarantees for the convergence should the condition $n \geq 2f + 3$ hold true. At any training round, for each model update δ_i, Krum takes the following steps: (a) computes the pairwise euclidean distance of $n - f - 2$ updates that are closest to δ_i, (b) computes the sum of squared distances between update δ_i and its closest $n - f - 2$ updates. Then, Krum chooses the model update with the lowest sum to update the parameters of the joint global model.

Coordinate-Wise Median. In Coordinate-Wise Median (CWM) aggregation rule (Yin et al., 2018), for each j_{th} model parameter, the j_{th} coordinate of received model updates are sorted, and their median is used to update the corresponding parameter of the global model.

Trimmed Mean. Trimmed Mean (TM) is a coordinate wise aggregation rule (Yin et al., 2018). for $\beta \in [0, \frac{1}{2})$, trimmed mean computes the j_{th} coordinate of aggregate of n model updates as follows: (a) it sorts the j_{th} coordinate of the n updates, (b) discards the largest and smallest β fraction of the sorted updates, and (c) takes the average of remaining $n(1 - 2\beta)$ updates as the aggregate for the j_{th} coordinate.

RFA. RFA (Pillutla et al., 2019) is a robust privacy-preserving aggregator which requires a secure averaging oracle. RFA aggregates local models by computing an approximate of the geometric median of their parameters using a variant of the smoothed version of Weiszfeld's algorithm (Weiszfeld, 1937). RFA appears to be tolerant to data poisoning attacks but can not offer byzantine tolerance as it still requires clients to compute aggregation weights according to the protocol. Relying on clients to follow a defensive protocol without a proper means to attest to the correctness of computations on the client-side cast doubts on the practicality of RFA. To the best of our knowledge, RFA is the only existing defense that is compatible with secure aggregation.

Norm Bounding. Norm Bounding (NB) is an aggregation rule proposed by (Sun et al., 2019), which appears to be robust against false-label backdoor attacks. In this aggregation rule, a norm constraint M is set for model updates submitted by clients to normalize the contribution of any individual participants. Norm bounding aggregates model updates as follows: (a) model updates with norms larger than the set threshold M are projected to the $l2$ ball of size M and then (b) all model updates are averaged to update the joint global model.

Differential Privacy. Differential Privacy (DP) originally was designed to establish a strong privacy guarantee for algorithms on aggregate databases, but it can also provide a defense against poisoning attacks (Ma et al., 2019; Dwork et al., 2006). Extending DP to FL ensures that any participant's contribution is bounded and therefore, the joint global model does not over-fit to any individual update. DP is applied in FL as follows (Kairouz et al., 2019): (a) server clips clients' model update by a norm M, (b) clipped updates are aggregated, then (c) a Gaussian noise is added to the resulted aggregate. DP has recently been explored and shown to be successful against false-label backdoor attacks in a study by (Sun et al., 2019).

C.1 IMPLEMENTATION DETAILS OF DEFENSES

As mentioned in §3, we compare the performance of contemporary defense such as Krum, Coordinate-wise Median (CWM), trimmed mean (TM), norm bounding (NB), differential privacy (DP), and RFA against naive and model replacement backdoor attacks on both Meta-FL and baseline FL.

Hyper parameter and implementation details of these techniques are provided below: **(1)** For Krum, to meet the convergence condition $n \geq 2f + 3$, we set $f = 6$. **(2)** In Trimmed mean, the parameter β is set to 0.20. **(3)** For RFA, the maximum iteration of the Weiszfeld algorithm and the smoothing factor is set to 10, and 10^{-6}, respectively. **(4)** In norm bounding defense, as the original work (Sun et al., 2019) did not provide a recipe to decide the norm threshold M, we developed our own approach to determine M. In our experiments, at each round, we set the norm threshold M to the norm of the smallest aggregand to ensure all aggregands will have an equal $l2$ norm before aggregation. In FL, as the global model converges, model updates (and therefore cohort aggregates) start to fade out and have smaller norms. Therefore, setting a constant norm threshold for all training rounds would not be effective, which is why we took a dynamic approach to decide M. **(5)** For differential privacy the hyper-parameter M is set similar to norm bounding and then a Gaussian noise $\mathcal{N}(0.0,\, 0.001^2)$ is added to the aggregate of updates (or cohort aggregates) before updating the global model.

D Background Knowledge

D.1 Federated Learning

Federated learning is a machine learning setting that enables millions of clients (mobile or edge devices) to jointly train a deep learning model using their private data without compromising their privacy. The training procedure in FL is orchestrated by a central server responsible for providing the shared global model to participants and aggregating their submitted model updates to generate the new global model. The key appeal of FL is that it does not require centralizing participating users' training data, which makes it ideal for privacy-sensitive tasks.

A standard FL setting consists of P participating clients. Each client i holds a shard of training data D_i which is private to the client and is never shared with the orchestration server. In each round t of FL, the central server randomly selects a set ζ^t of c clients, and broadcasts the current global model G^t to them. Selected set of clients ζ_t is referred to as *training cohort* of round t. Each client i in the training cohort locally and independently trains the joint model G^t using Stochastic Gradient Descent (SGD) optimization algorithm for E epochs on its local training data D_i to obtain a new local model L_i^{t+1}, and submits the difference $L_i^{t+1} - G^t$ as its model update to the central server. Next, the central server averages model updates submitted by clients in the training cohort and updates the shared global model using its learning rate η to obtain the new global model G^{t+1}, as shown in Equation 2. Model training resumes until the global model converges to acceptable performance, or certain training rounds are completed.

$$G^{t+1} = G^t + \frac{\eta}{n} \sum_{i=i}^{n} (L_i^{t+1} - G^t) \tag{2}$$

D.2 Secure Aggregation

Secure Aggregation (SecAgg) (Bonawitz et al., 2017) is a secure multi-party computation protocol that can reveal the sum of submitted model updates to the server (or aggregator) while keeping each individual update uninspectable at all time. Secure Aggregation consists of three phases, *preparation*, *commitment* and *finalization* (Bonawitz et al., 2019). In the preparation phase, shared secrets are established between the central server and participating clients. Model update from clients who drop out during the preparation phase will not be included in the aggregate. Next, in the commitment phase, each device uploads a cryptographically masked model update to the server, and the server computes the sum of the submitted mask updates. Only clients that successfully commit their masked model updates will contribute to the final aggregate. Lastly, in the finalization phase, committed clients reveal sufficient cryptographic secrets to allow the server to unmask the aggregated model update.

E THREAT MODEL

In this section, we present the objectives, capabilities, and schemes of backdoor attackers that are commonly used in prior studies. In other words, our proposed Meta-FL framework does not make any additional assumptions.

Attacker's Objective. Similar to prior arts such as (Xie et al., 2020; Bagdasaryan et al., 2020), we consider an adversary whose goal is to cause misclassifications to a targeted label T for inputs embedded with an attacker-chosen trigger. As opposed to Byzantine attacks (Blanchard et al., 2017), whose purpose is to convergence the learning algorithm to a sub-optimal or utterly ineffective model, the adversary's goal in backdoor attacks is to ensure that the joint global model achieves high accuracy on both the backdoor sub-task and the primary learning task at hand.

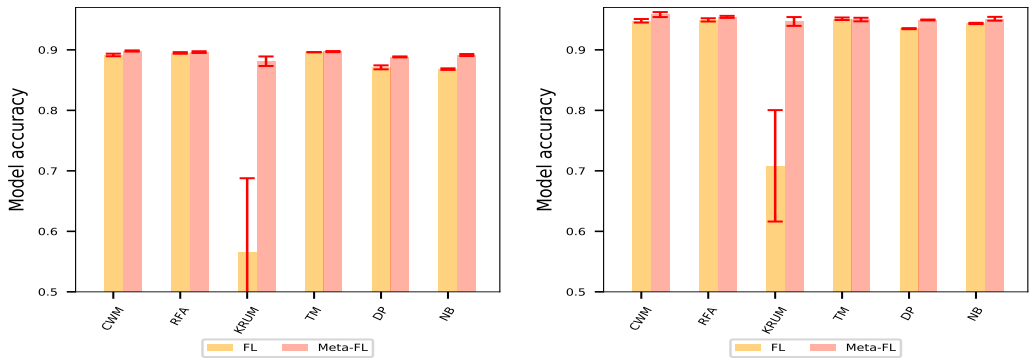
Attacker's Capability. We make the following assumptions about the attacker's capabilities: (a) We assume the attacker controls a number of participants, which are referred to as *sybils* in the literature of distributed learning. Sybils are either malicious clients which are injected into FL system or benign clients whose FL training software has been compromised by the adversary, (b) following Kerckhoffs's theory (Shannon, 1949), we assume a strong attacker who has complete control over local data and training procedure of all its Sybils. The attacker can modify training procedure's hyperparameters and is capable of modifying model updates before submitting them to the central server, (c) adversary is not capable of compromising the central server or influencing other benign clients, and more importantly, does not have access to benign clients' local model, training data and submitted updates.

Attack scheme. In our evaluations, we consider two backdoor attack schemes which are referred to as "Naive" and "Model Replacement" in literature (Bagdasaryan et al., 2020). In both schemes, adversaries train their local model with a mixture of clean and backdoored data, and model updates are computed as the difference in the parameters of the backdoored local model and the shared global model. In the naive approach, the adversary submits the computed model update. While in model replacement attack, the model update is scaled using a scaling factor to cancel the contribution of other benign clients and increase the impact of the adversarial update on the joint global model. A carefully chosen scaling factor for adversarial updates can guarantee the replacement of the joint global model with the adversary's backdoored local model.

F META-FL VS BASELINE FL [DETAILED VERSION]

F.1 UTILITY

In this section, we compare the utility of Meta-FL against baseline setting in terms of model accuracy. In this experiment, we evaluate the utility of baseline and Meta-FL frameworks across various FL configurations and aggregation rules. Note that for a fair comparison, we make sure the number of clients participating in each round of model training is equal across both frameworks. Figure 4 reports the test accuracy of models trained in Meta-FL and baseline settings deploying different defenses and aggregation rules. As reflected, federated training with Meta-FL results in more accurate models compared to baseline setting. All defenses and aggregation rules offer better utility in our framework. Even Krum aggregation rules which has been known to cause a large drop in performance of the learned model in baseline FL (Bagdasaryan et al., 2020; Bhagoji et al., 2019) can train models with comparable performances in Meta-FL.



(a) SVHN (b) GTSRB

Figure 4: Comparing utility of Meta-FL against baseline FL in terms of model accuracy.

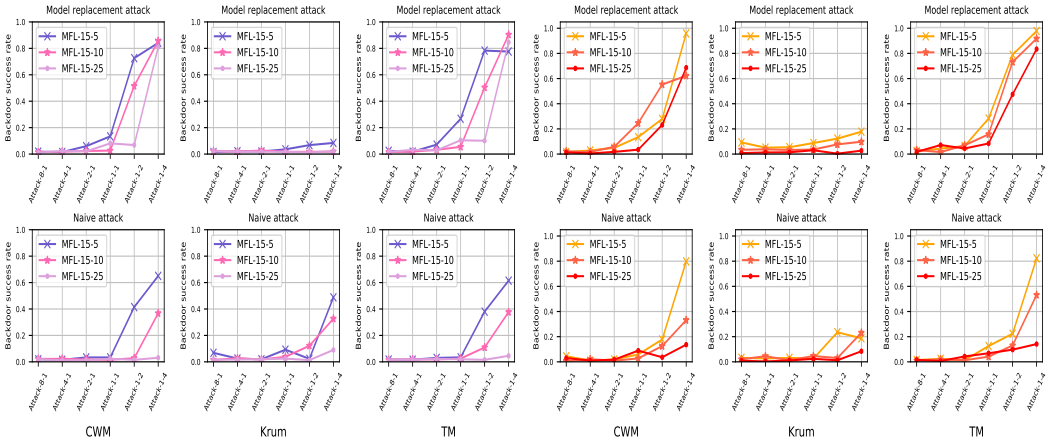
F.2 ROBUSTNESS

In this section, we systematically compare the capabilities of contemporary defenses against backdoor attacks in both baseline and meta federated learning. Our empirical evaluation in this section shows that all defenses benefit from the advantages discussed in §2 and offer better robustness in our framework Meta-FL.

Figures 1 and 2 report performance of contemporary defenses against backdoor attacks on GTSRB and SVHN benchmarks, respectively. We experiment with several attack scenarios to systematically evaluate the performance of each defense against adversaries with a wide range of resources at hand. As we move along the attack scenarios denoted on the horizontal axis of diagrams in Figures 2 and 1, the adversary becomes more and more powerful and appears more frequently with more sybils at each round.

For a fair evaluation of contemporary defense across Meta-FL and baseline FL, we make sure the defender faces similar challenges in both frameworks. Throughout our experiments in this section, we set the number of training cohorts in Meta-FL equal to the number of selected clients in baseline FL to ensure that server sees the same number of "aggregands" (client updates in baseline FL and cohort aggregates in Meta-FL) across both cases. Moreover, the way our attack scenarios are defined ensures that the same number of aggregands are adversarial across both frameworks.

Across both Meta-FL and baseline FL frameworks, the scaling factor for model replacement attack is set equal to the size of training cohorts to ensure that submissions from adversarial clients survive the averaging procedure and overpower the aggregate of their corresponding cohort. For attack scenarios in which multiple sybils appear in the same round, we assume they coordinate and divide the scaling factor among themselves evenly.



(a) SVHN (b) GTSRB

Figure 5: Effect of size of training cohorts on efficacy of CWM, Krum and TM against backdoor attacks.

Figures 1 and 2 show that *Meta-FL puts all defense at an advantage in mitigating against backdoor attacks.* Attack success rate of both the naive and model replacement approach in Meta-FL (solid lines) is lower than in baseline FL (dashed lines) when the same defense is in place across both frameworks. Therefore, our empirical evaluations show that existing defenses are more robust to backdoor attacks in Meta-FL compared to baseline FL across.

While Meta-FL enhances the robustness of all 6 methods, we observe that Krum benefits the most from our framework. We believe that lower variance on cohort aggregates aids Krum to effectively separate benign and malicious updates. We note that server can further decrease the variance of cohort aggregates along each coordinate by increasing the size of training cohorts, As discussed in §2, and improve robustness of Krum aggregation rule.

Moreover, other methods such as coordinate-wise median and trimmed mean which are anomaly detection-based defenses can also benefit from lower variations on cohort aggregate. Perhaps the most important principle in detecting outliers is defining the distribution of ordinary observations, which can be easier should observations exhibit low variations. Figure 5 shows the results for experiments in which we evaluate the performance of Krum, CWM, and TM across Meta-FL frameworks with increasingly larger training cohorts. For this experiment, we set the number of cohorts to 15 and varied cohort size between 5, 10 and 15. As reflected in Figure 5, increasing the size of training cohorts improves the robustness of these techniques across all scenarios, especially for scenarios in which the adversary appears more frequently with more sybils.

Although defenses such as RFA, differential privacy, and norm bounding appear to be robust against poisoning attacks (Sun et al., 2019; Pillutla et al., 2019), our empirical evaluations show that they are not effective against backdoor attacks, specifically model replacement attacks. In poisoning attacks, the adversarial sub-task, which is misclassification of unmodified data samples (e.g. classifying certain images of digit 1 as digit 7), is in direct contradiction with the primary learning task. Therefore, poisoning updates (or aggregates) face direct opposition from submissions of benign clients, which makes it harder for the adversary to succeed. However, for the case of backdoor attacks, the adversary's goal for the model is to learn the causal relation between the presence of an attacker's chosen trigger and certain model output which does not require the model to learn any knowledge contradicting the primary learning task. Therefore, backdoor attacks tend to be stealthier compared to poisoning attacks, and defenses that have shown resilience against poisoning attacks might fall short against backdoor attacks.