FLARE: Defending Federated Learning against Model Poisoning Attacks via Latent Space Representations

Ning Wang
Virginia Tech
Blacksburg, VA, United States
ning18@vt.edu

Yang Xiao
Virginia Tech
Blacksburg, VA, United States
xiaoy@vt.edu

Yimin Chen
University of Massachusetts Lowell
Lowell, MA, United States
ian_chen@uml.edu

Yang Hu
Virginia Tech
Blacksburg, VA, United States
yanghu@vt.edu

Wenjing Lou
Virginia Tech
Blacksburg, VA, United States
wjlou@vt.edu

Y. Thomas Hou
Virginia Tech
Blacksburg, VA, United States
thou@vt.edu

ABSTRACT

Federated learning (FL) has been shown vulnerable to a new class of adversarial attacks, known as *model poisoning attacks (MPA)*, where one or more malicious clients try to poison the global model by sending carefully crafted local model updates to the central parameter server. Existing defenses that have been fixated on analyzing model parameters show limited effectiveness in detecting such carefully crafted poisonous models. In this work, we propose FLARE, a robust model aggregation mechanism for FL, which is resilient against state-of-the-art MPAs. Instead of solely depending on model parameters, FLARE leverages the *penultimate layer representations (PLRs)* of the model for characterizing the adversarial influence on each local model update. PLRs demonstrate a better capability to differentiate malicious models from benign ones than model parameter-based solutions. We further propose a trust evaluation method that estimates a trust score for each model update based on pairwise PLR discrepancies among all model updates. Under the assumption that honest clients make up the majority, FLARE assigns a trust score to each model update in a way that those far from the benign cluster are assigned low scores. FLARE then aggregates the model updates weighted by their trust scores and finally updates the global model. Extensive experimental results demonstrate the effectiveness of FLARE in defending FL against various MPAs, including semantic backdoor attacks, trojan backdoor attacks, and untargeted attacks, and safeguarding the accuracy of FL.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

federated learning; model poisoning attack; defense

ASIA CCS '22, May 30–June 3, 2022, Nagasaki, Japan

ACM ISBN 978-1-4503-9140-5/22/05.
https://doi.org/10.1145/3488932.3517395

ACM Reference Format:
Ning Wang, Yang Xiao, Yimin Chen, Yang Hu, Wenjing Lou, and Y. Thomas Hou. 2022. FLARE: Defending Federated Learning against Model Poisoning Attacks via Latent Space Representations. In *Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22), May 30–June 3, 2022, Nagasaki, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3488932.3517395

1 INTRODUCTION

Machine learning (ML) is changing the ways people live and do business in every sector of our society. The success of ML, especially deep learning (DL), relies on the availability of powerful computers and massive amount of training data. However, learning systems that require all the data to be fed into a learning model running on a central server pose serious privacy concerns. For example, the transmission of health data across certain organizational boundaries may violate security and privacy rules such as those imposed by the Health Insurance Portability and Accountability Act (HIPAA[1]). Federated learning (FL) [16, 18, 22], which enables a group of intelligent agents to jointly learn a model while keeping their private data at their local devices, emerges as a promising new learning framework to address client data privacy problems.

FL has been applied in many popular applications, such as next-word prediction on Android Gboard by Google [14] and credit risk control by WeBank [31]. In an FL system, a large number of distributed clients cooperatively contribute to the learning process by uploading the gradients of their local models (or model weights) to the *parameter server* (*PS*) through multiple iterations without sharing the raw data at the clients. At the beginning of an FL task, *PS* initializes a global model. In each learning iteration, *PS* distributes the current global model parameters to selected clients. Each selected client continues to train the received model with its local data independently by following a predefined learning protocol. At the end of each learning iteration, *PS* collects and aggregates updates from clients using a gradient aggregation rule such as `FedAvg` [22]. *PS* then updates its global model and after multiple iterations *PS* outputs the final global model.

Despite many salient features of FL and its tremendous success in many applications, it has been shown recently that FL is vulnerable to model poisoning attacks (MPAs) [4, 8, 10, 13, 29]. In an MPA, the attacker (i.e., a malicious client) manipulates or crafts its model

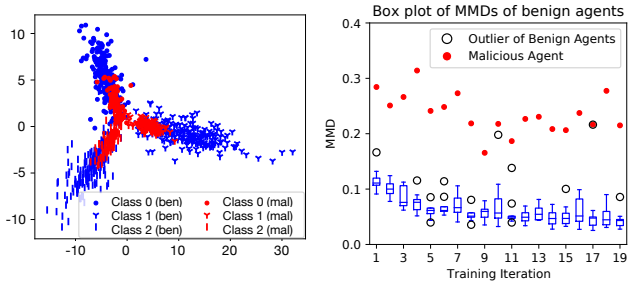
[1] https://www.hhs.gov/hipaa/index.html

parameters sent to the *PS* in the aim of corrupting the global model by either increasing the prediction error (untargeted attacks) [10] or controlling the prediction on targeted inputs (backdoor attacks) [3, 4]. It is shown in [4] that even a single malicious attacker could deteriorate the global model accuracy and succeed in controlling the model output on chosen input data.

A potential countermeasure to MPAs is Byzantine resilient aggregation rules (BRARs) [5, 9, 34], which enable *PS* to learn an accurate global model when a bounded number of clients are malicious (i.e., Byzantine). Compared to straightforward aggregation rules that linearly combine the model updates (e.g., FedAvg [22]), BRARs (e.g., `Krum` [5]) seek to provide statistical methods that are not abused by Byzantine values. To this end, BRARs leverage outlier-robust measures [15], e.g., median, trimmed estimator, to compute the center of updates despite the presence of Byzantine updates. Another line of defenses [20, 27] resorts to use anomaly detection methods to detect malicious local model updates and excludes them from the aggregation. MPAs have seen an increase in stealthiness and sophistication. The state-of-the-art MPAs [4, 10] can craft malicious model updates very similar to benign ones, breaking existing BRARs. Both the BRARs and ML-based defenses explore the model parameter space for detecting anomalous updates; they nonetheless show limited effectiveness in defending against the state-of-the-art MPAs [4, 10]. Many ML-based defenses [11, 20, 27] also need to collect a dataset of labeled benign and malicious models beforehand.

In this paper, we tackle the MPA challenge of FL through a new angle—the latent space representation of a model. We first make an important observation that even though the poisoned model parameters are very close to those of benign models, their representations in the latent space, provided an auxiliary input dataset, tend to diverge from those of benign models. Specifically, we target the *penultimate layer representation (PLR)* vector in the latent space and plot the PLRs of both attack-free models and poisoned models in Figure 1(a). It shows that the clean/benign PLRs follow the same distribution while the poisoned/malicious PLRs follow a different one. We made such observation consistently across different datasets and different neural network architectures. Besides the visual differences, to obtain quantifiable discrepancy, FLARE measures the distance (i.e., maximum mean discrepancy (MMD) [12]) between the PLRs of any two models. The average MMD scores of both poisoned models and clean models are illustrated in Figure 1(b), which confirms that PLR is a highly differentiating feature for poisonous models.

Based on the above observation, we propose FLARE (**F**ederated learning+**LA**tent-space **RE**presentations) to protect FL systems against state-of-the-art MPAs. FLARE features a novel methodology to estimate the trust score of a local model update by exploiting the similarity between its PLR to the PLR of others. Compared to defenses that only look into the model parameters, the PLR-based trust estimation enables FLARE to prevail in defending against carefully crafted malicious model updates. To estimate the trust score, FLARE computes a PLR sequence for each local model, which takes a very small auxiliary data at the *PS*. FLARE then exploits PLRs to distinguish malicious model updates from benign ones. Under the assumption that malicious clients are fewer than honest clients, FLARE assigns a trust score to each model update based on the pairwise PLR discrepancies among all model updates, in



(a) 3-class clusters (b) MMD distribution

Figure 1: A motivating example for our PLR approach using the fMNIST dataset [32]. (a) PLRs of 3 classes projected in a 2D space: benign models' PLRs are in blue and poisoned models' PLRs in red. (b) Averaged maximum mean discrepancy (MMD) between a model's PLRs and other models' PLRs. Box is for benign models, while red dots denote malicious model.

that those farther from the benign distribution are assigned lower scores. Finally, we employ a soft decision regime that aggregates model updates weighted by their trust scores. It is worth noting that FLARE performs trust score estimation based on the most recently received model parameters in each federated learning iteration, and it does not require collecting a dataset of model parameters beforehand, which yields efficiency advantage compared to existing ML-based defenses [11, 20, 27].

Contributions of this paper are summarized as follows:

- We propose FLARE, a novel detection and aggregation algorithm for FL to defend against state-of-the-art MPAs. Based on the key observation that PLRs of poisoned models tend to diverge from those of benign models, FLARE utilizes PLR for evaluating the trust score of a model update in FL. Based on the MMD of different local models' PLRs, FLARE features a trust estimation mechanism that assigns a trust score to each client in every learning iteration. FLARE minimizes the impact of MPAs by aggregating model updates weighted by their client trust scores.
- Through theoretical analysis, we provide an Euclidean-distance-based interpretation on PLRs of deep neural network (DNN), justifying PLR as a promising measure to estimate the trust score of a model update.
- Extensive experimental results demonstrate the effectiveness of FLARE for defending against state-of-the-art MPAs. FLARE outperforms existing defenses in terms of decreasing the attack success rate of MPAs. FLARE achieves consistent performance across various attack methods, and datasets, demonstrating the generality of the approach.

2 BACKGROUND AND RELATED WORK

Federated learning (FL), in a nutshell, allows a group of distributed clients to contribute their locally computed model parameter updates to the global model at the parameter server. The parameter server is responsible for distributing the initial model, collecting model parameter updates from agents, aggregating them through a

certain aggregation rule, and adding the result to the global model. Eyeing on this FL paradigm, a class of stealthy attacks named model poisoning attacks (MPAs) have been demonstrated to be a significant threat to the security of FL systems [3, 4, 10, 24]. In an MPA, a compromised local agent attempts to corrupt the training process of FL by providing the parameter server carefully manipulated model parameters in each training iteration, in the aim of gradually degrading the FL model efficacy without being detected.

To protect the global model from malicious local updates in FL systems, BRARs were proposed in the literature, exampled by `Krum` [5], `Coomed`, `Trimmed Mean` [34], and `Bulyan` [9]. BRARs tackle the Byzantine attack/failure scenario in FL where a client does not follow the predefined learning protocol and sends arbitrary model updates to the *PS*. Technically, BRARs can bound the gap between the aggregated gradient and the true mean (i.e., without Byzantine clients) to a small value. Based on this feature, BRARs can partially address the MPA threat by preventing or downgrading the impact of some malicious model updates. Below we briefly introduce four state-of-the-art BRARs. `Krum` [5] selects one of n received updates $\{\delta_1, ..., \delta_n\}$ whose distance to the all the remaining update is the smallest. `Coomed` [34] selects the coordinate-wise median of n received updates as the final result. `Trimmed Mean` [34] first excludes the largest k values and the smallest k values in each coordinate. Then it calculates the average value of the remaining $(n - 2k)$ items. `Bulyan` [9] is a combination of `Krum` and `Trimmed Mean`. `Bulyan` firstly recursively applies `Krum` to select $(n - 2k)$ updates out of the total n updates. Then it applies `Trimmed Mean` to the selected $(n - 2k)$ updates to obtain the final result. We will use these BRARs for comparative analysis and evaluation.

Besides BRARs, a number of anomaly detection mechanisms are proposed to detect malicious local model updates. Shen et al. [27] proposed *Auror* to protect FL from malicious updates by filtering out-of-distribution parameters from the received model parameters; Fung et al. [11] proposed *FoolsGold* to identify poisoning Sybils based on the model similarity of client updates. Li et al. [20] proposed a spectral-anomaly-detection-based framework that detects the abnormal model updates based on their low-dimensional embeddings. Zhao et al. [35] proposed PDGAN for detecting poisoned models. PDGAN reconstructs training data from model updates and audits the accuracy for each participant model by using the generated data, and removes clients with accuracy lower than a predefined threshold. [1] uses a set of validating clients to determine if the (global) model-update derived in that round has been subject to a poisoning injection. That is, clients validate the global model on their local data, and vote for accepting or rejecting the model through a feedback loop. [7] proves that majority vote mechanism with ensemble federated learning is secure against MPA. The most relevant work to ours is FLTrust [6]. FLTrust bootstraps a trust score for each client based on its directional deviation from server model update and computes the average of the local model updates weighted by their trust scores as a global model update.

In the meantime, MPAs have seen an increase in stealthiness and sophistication. The backdoor MPAs proposed by Bhagoji et al. [4] incorporate a penalty on the distance between the crafted model parameters and the benign model parameters into its optimization objective. Bagdasaryan et al. [3] developed a generic constrain-and-scale technique that incorporates the evasion of defenses into

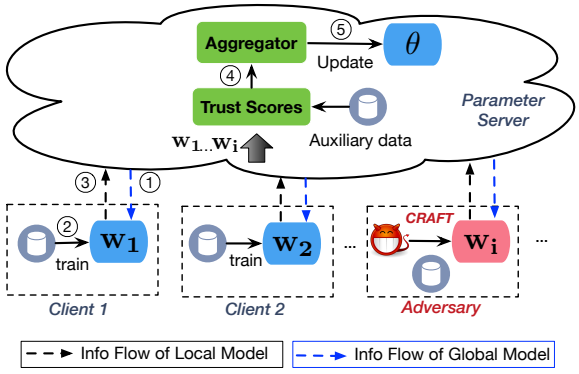


Figure 2: Federated learning system model.

the attacker's loss function during training. Similar techniques have been adopted in later works [30, 33] to achieve evasion of defenses. Meanwhile, Fang et al. [10] proposed untargeted MPAs to degrade the overall accuracy of the FL system by deviating the crafted model parameters from the true gradient direction. These MPAs [3, 10] have demonstrated their capability in evading existing defenses, e.g., `Krum`, `Trimmed Mean`, *Auror* and *FoolsGold*. [3] shows that an attacker is able to craft a malicious model satisfying that the Euclidean distance between the crafted model and any benign model is comparable or even less than the Euclidean distance among different benign models. Moreover, this crafted model can still misclassify an input to a target label. This attack makes the defenses by exploring Euclidean distance of model parameters useless and leads us to reconsider the defense for the MPAs.

We observe that most of the malicious model detection mechanisms [11, 20, 27], BRARs [5, 9, 34], and client credibility aggregation mechanisms (e.g., FLTrust [6] and [2, 21]) build their defense by directly analyzing the model updates from agents in the model parameter space. We also observe that due to the high dimensionality of the FL model as well as the non-smooth loss function, two models that are seemingly close in the parameter space may have dramatically different loss function. These defenses are likely to make miss detection on a malicious model that is carefully crafted to be similar to benign models in the parameter space. Based on this key insight, we propose to detect malicious local models by analyzing the latent-space features of models.

3 SYSTEM MODEL

3.1 Federated Learning with Trust Scores

We consider a typical FL network with one parameter server *PS* and n participating clients $\{C_i\}|_{i\in[n]}$ (we define $[n] := \{1, 2, ..., n\}$). The definition of frequently used symbols are shown in Table 1. Each client manages a local model (e.g., a neural network). At *PS*, the model weights of C_i are $\mathbf{w}_i \in \mathcal{W} \subseteq \mathbb{R}^d$, wherein $\mathcal{W}$ is the parameter space and d is the presumed model dimensionality. The global model parameter is denoted by $\theta \in \mathcal{W}$. We denote the model update from C_i as $\delta_i = \mathbf{w}_i - \theta$. Moreover, *PS* maintains a vector of trust scores for all clients, denoted $\{S_i\}_{i\in[n]}$. During normal operation, an FL task executes in iterations with *PS* acting as the model distributor and aggregator at the cloud side.

Figure 2 illustrates the FL system model. At the system onset, *PS* initializes θ. Then each training iteration works as follows: ① *PS* first selects multiple clients and distributes θ to them; ② each of the selected clients, say C_i, initializes $\mathbf{w}_i = \theta$ and trains the model with its local data; ③ after the local training terminates, C_i provides its model update δ_i to *PS*; ④ *PS* uses the local model updates provided by the clients to compute a trust score S_i for every client C_i; ⑤ finally, *PS* aggregates local model updates weighted by their trust scores and updates the global model by: $\theta \leftarrow \theta + \sum_i S_i \delta_i$. At the end of the FL task, *PS* outputs the final global model.

We remark that the trust scores do not exist in the original FL formulation. As we show later, they allow our system FLARE to be responsive to malicious model updates before the aggregation step. We also assume *PS* has an auxiliary dataset $\mathcal{D} = \{x_i\}_{i\in[m]}$ containing a small number of records (e.g., $m = 10$), which will be used for PLR-based trust score evaluation in step ④. $\mathcal{D}$ can be obtained as long as we have one or more trusted clients in the FL system, who are willing to contribute to the system's security.

3.2 Threat Model

We assume the population of malicious clients is less than $0.5n$, in line with prior work [5, 6, 9, 34]. Meanwhile, every malicious client is a white-box adversary and can mount MPAs on the system, following from the state-of-the-art MPAs [4, 10].

White-Box Adversary. Being a valid FL client, the attacker has access to both the global model parameters and the model updates of other clients. Typically, the attacker estimates the model updates of other clients using a dummy model trained on its own clean data. Compared to a white-box attacker, a black-box attacker would only have access to the global model parameters. We opt for the challenging white-box adversary model to demonstrate the strength and effectiveness of our proposed defense.

Model Poisoning Attacks. According to the adversary goal, there are two types of MPAs: untargeted attacks and backdoor attacks. For an untargeted attack, the attacker aims to degrade the overall model accuracy. For a backdoor attack, the attacker aims to control the predictions on chosen input data records without degrading the overall prediction performance on other input data records. In specific, we use the two untargeted attacks in [10], which deviate the global model toward the opposite of the attack-free direction. We use the two backdoor attacks in [4], in which the adversary crafts malicious local models so as to inject a backdoor/trigger into the global model. The adversary maintains its stealth by decreasing the distance between the crafted model parameters and benign model parameters.

4 PENULTIMATE LAYER REPRESENTATION

We present the motivation, theory, and outlook for using penultimate layer representations (PLRs) to defend against MPAs.

4.1 PLR Basics

We use a convolutional neural network (CNN) for instance. Consider a CNN $f : \mathbb{R}^{d1\times d2\times d3} \rightarrow \mathbb{R}^c$, mapping points $x \in \mathbb{R}^{d1\times d2\times d3}$ to a c-dimensional probability vector $\mathbf{q} \in \mathbb{R}^c$, where c is the number of classes. We consider an image input and use $d1$, $d2$, and $d3$ to represent an image's width, height, and number of channels. Let the last layer of the network be a softmax layer. The mapping function from the input to the penultimate layer (i.e., the layer before the last layer) is denoted by $g : \mathbb{R}^{d1\times d2\times d3} \rightarrow \mathbb{R}^o$. The output of function g is a PLR which is denoted by $\mathbf{r} \in \mathbb{R}^o$. We use σ to denote the mapping function from PLR to the output probability vector, $\sigma : \mathbf{r} \in \mathbb{R}^o \rightarrow \mathbf{q} \in \mathbb{R}^c$. The mapping functions are shown in Figure 3.

Table 1: Symbol definition.

Symbol	Definition
PS	parameter server
C_i	the i-th client
n	the number of clients
$\mathbf{w}_i$	the model parameters of C_i
θ	the global model parameters at *PS*
δ_i	the model parameter update of C_i
$\mathcal{D}$	the auxiliary dataset, $\lvert\mathcal{D}\rvert = m$
m	the number of data points in $\mathcal{D}$
c	the number of classes of the data
$\mathbf{x}$	the input image, $\mathbf{x} \in \mathbb{R}^{d_1\times d_2\times d_3}$
$\mathbf{r}$	the penultimate layer representation, $\mathbf{r} \in \mathbb{R}^o$
$\mathbf{q}$	the confidence vector $\mathbf{q} \in \mathbb{R}^c$
R	a sequence of $\mathbf{r}$, $R = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_m\}$
f	the mapping function of $\mathbf{x} \rightarrow \mathbf{q}$
g	the mapping function of $\mathbf{x} \rightarrow \mathbf{r}$
σ	the mapping function of $\mathbf{r} \rightarrow \mathbf{q}$
ω_i	model weights from the penultimate layer to the i-th output neuron
Ω	$\Omega = \{\omega_1, \omega_2, ..., \omega_c\}$
ct_i	the count that C_i is selected as others' nearest neighbor
S_i	the trust score of C_i

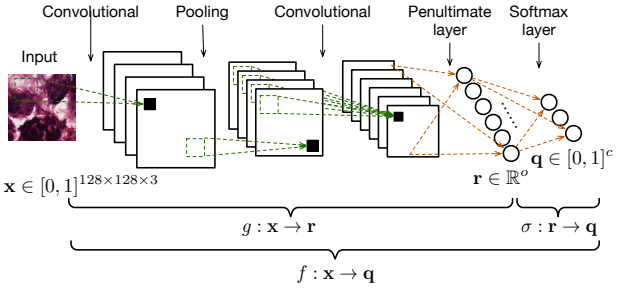


Figure 3: The convolutional neural network architecture showing mapping functions f, g and σ.

4.2 Power of PLR in Separating MPAs

Next we show that PLR exhibits highly differentiating power in detecting malicious models crafted by the advanced attacks. This is in contrast to solely looking at the model parameters.

The prediction of a c-class classifier on an input is a probability vector $\mathbf{q} = [q_1, q_2, ..., q_c]$, where q_k represents the likelihood the model assigns label k to the input and $\sum_{k=1}^{c} q_k = 1$. We use $\Omega = [\omega_1, , \omega_2, ..., \omega_c]$ to represent the weight connecting the penultimate layer to the last layer where $\omega_k \in \mathbb{R}^o$ denotes the weights connecting to the k-th neuron of the output (i.e., softmax) layer.

According to the softmax function, q_k is calculated as

$$q_k = \frac{\exp(\mathbf{r}^T \omega_k)}{\sum_{i=1}^{c} \exp(\mathbf{r}^T \omega_i)}. \quad (1)$$

We interpret the output probability using the Euclidean distance between PLR and templates ω_k.

Proposition 1. *In a c-class NN classifier where the last two layers are fully connected and the last layer is a softmax layer, the output probabilities of any two class k and l ($\forall k, l \in [c]$ and $k \neq l$) satisfy that $q_k > q_l$ if*

$$\|\mathbf{r} - \omega_l\|_2 - \|\mathbf{r} - \omega_k\|_2 \geq C_{kl}, \quad (2)$$

where $\mathbf{r}$ represents the PLR of an input data record and ω_k is the weights connecting to the k-th neuron of the output layer. $||\mathbf{r} - \omega_k||_2$ denotes the Euclidean distance between $\mathbf{r}$ and template ω_k, i.e., $||\mathbf{r} - \omega_k||_2 = \mathbf{r}^T\mathbf{r} - 2\mathbf{r}^{\mathbf{T}}\omega_k + \omega_k^T\omega_k$. C_{kl} is a constant and $C_{kl} = \omega_l^T\omega_l - \omega_k^T\omega_k$.

Proposition 1 (see proof in Appendix) implies that the smaller the distance between $\mathbf{r}$ and the template ω_k (when the distance between $\mathbf{r}$ and other templates is fixed), the larger the likelihood that $\mathbf{r}$ is classified as class k. Here, we can regard the template ω_k as the cluster center of class k. Classification can be determined by comparing a target PLR to all the c templates. For each pair of classes, e.g., k and l, the input is more likely to be class k if Eq. (2) is satisfied or class l otherwise. Based on Proposition 1, we hypothesize that the PLRS of inputs belonging to one specific class exhibit a relatively small Euclidean distance to the corresponding template, possibly resulting in a consistent pattern (i.e., a cluster centered at the template). We then analyze how the distortion in PLR (i.e., $\|\mathbf{r}_1 - \mathbf{r}_2\|_2$) transforms to the final output probability vector.

Proposition 2. *The mapping function $\sigma : \mathbf{r} \in \mathbb{R}^o \rightarrow \mathbf{q} \in \mathbb{R}^c$ maps a PLR to a probability vector as discussed above. For any two PLRs $\mathbf{r}_1, \mathbf{r}_2$, we have*

$$\|\mathbf{q}^1 - \mathbf{q}^2\|_2 \leq \|\Omega\|_2 \|\mathbf{r}_1 - \mathbf{r}_2\|_2, \quad (3)$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the PLR of two input x_1 and x_2 respectively. $\mathbf{q}^1$ and ($\mathbf{q}^2$) are the output probability vector for input x_1 and x_2 respectively.

Proposition 2 (see proof in Appendix) implies that a difference in the PLR space will transform to the difference of the corresponding output probability. The output probability will be very similar if $\|\mathbf{r}_1 - \mathbf{r}_2\|_2$ is small enough.

Furthermore, we consider two local models in FL. The weights Ω of the two local models should be very similar since the local models begin their training with the same initialization and only perform several training steps. The two models transform the same input x (belonging to class c) into PLR $\mathbf{r}_1$ and PLR $\mathbf{r}_2$ respectively. $\mathbf{r}_1$ and $\mathbf{r}_2$ should stay within a certain distance if the two models would give a similar prediction confidence vector on x. We further demonstrate our hypothesis by visualizing the PLRs of benign models in Figure 4. Here we assume ten participating clients in an FL system, including nine benign clients and one attacker. The attacker manipulates its model parameters by implementing a well-known backdoor MPA [4]. We collect the ten local models and calculate a PLR for each

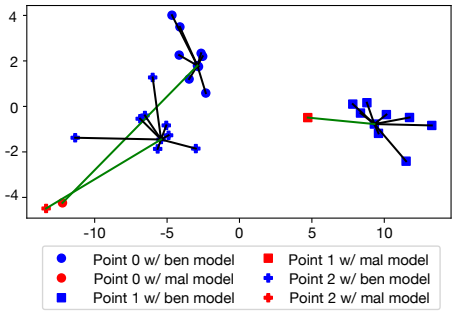


Figure 4: PLRs and inter-distance between PLRs in the 2-D space of the Kather dataset [17]. There are three types of markers (i.e., circle, plus, and square), each type of maker corresponds to one input data points. There are ten items of each type of marker, representing ten versions of PLRs for one input data. Nine (in blue) are from nine benign local models and one (in red) is from a malicious model.

local model using the same data point. We plot the ten versions of PLR in a 2-D space. Ten versions of PLR are presented in Figure 13. We can see that all the benign PLRs stays close to each other while the malicious one exhibits a more significant distance from the benign ones. We show three examples (three input data points), and we get a similar observation among all three examples.

4.3 Visualizing PLRs Distribution

Based on the above theoretical analysis and empirical results, we find that the poisoned model produces a PLR that is relatively far from the cluster of benign PLRs. We further plot more data points from the same class to see the distribution of PLRs. In Figure 6, we plot the PLRs of both malicious models and benign models under backdoor attacks. We observe that the PLRs of benign models follow a distribution while the PLRs of malicious models deviate from it. We present the results under untargeted attacks and achieve consistent result (see Figure 10). All the results further confirm that PLR is a promising feature for detecting poisoned models. The key reason for distribution deviation is because the PLR distance among benign models is smaller than the PLR distance between benign models and malicious models.

The visualization procedure used to illustrate the PLRs is as follows [23]: 1) randomly select three classes from all classes; 2) compute the orthonormal basis of the hyperplane on which the templates of the selected class reside; 3) project the PLRs of data records from the three classes to the hyperplane (i.e., calculating the inner product of the PLRs and the orthonormal basis); and 4) reduce the dimensionality of the results in Step 3) to 2-D space by applying PCA. Finally, we can plot the PLRs in 2-D space.

5 FLARE: DEFENDING AGAINST MPAS

5.1 Overview of FLARE

Based on the observation and theoretical analysis of PLRs, we design our system—FLARE. The overview of FLARE is shown in Figure 5. Our design is compatible with the general FL system. One

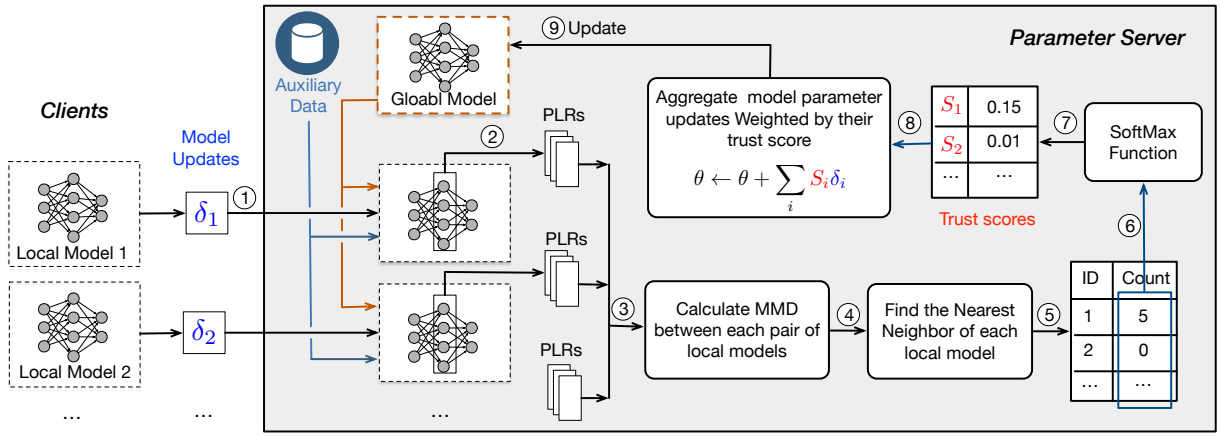


Figure 5: FLARE design. In each FL iteration, local clients submit their local model updates $\{\delta_i\}$ to the *PS*. *PS* calculates PLRs for each local model using the same auxiliary dataset, then computes the nearest neighbor of each local model based the MMD of PLRs. The count of being selected as a nearest neighbor are used to estimate the trust score S_i of each local model. *PS* aggregates the model updates weighted by their trust scores and uses the result to update the global model.

practitioner can easily apply FLARE to an FL system by adding a trust score estimation module. As shown in Figure 5, local clients firstly submit their local model updates δ_i to the *PS*. *PS* calculate PLRs using an auxiliary dataset for each local model. Next, FLARE computes the nearest neighbors of each local model based on the Maximum Mean Discrepancy (MMD) of PLRs. A client's count to be selected as other clients' nearest neighbor is used to estimate its trust score S_i. Finally, *PS* aggregates model updates weighted by their trust scores and use it to update the global model. The workflow of FLARE is also shown in Algorithm 1.

5.2 Detailed Design

FLARE features two differences at *PS* comparing to the traditional FL system: 1) At the very beginning, *PS* initializes the global model by training with the auxiliary dataset $\mathcal{D}$ instead of performing random initialization. This procedure can accelerate the training process. It also helps clients make correct predictions on $\mathcal{D}$, which increases the probability that the benign models' PLR follow one distribution. 2) At the aggregation stage of each learning iteration, *PS* estimate a trust score for every model update and aggregate them by their trust scores. Next we elaborate on FLARE's aggregation scheme, mainly how to estimate the trust scores.

Firstly, *PS* computes PLRs for each local model $\mathbf{w}_j$ using $\mathcal{D} = \{\mathbf{x}_i\}_{i\in[m]}$. The mapping function of the model with weight $\mathbf{w}_j$ is represented by $g_{\mathbf{w_j}}: \mathbf{x} \in \mathbb{R}^{d1\times d2\times d3} \rightarrow \mathbf{r} \in \mathbb{R}^o$ where $d1$, $d2$, and $d3$ represent the width, height, and channels of an image input and o represents the dimensionality of a PLR. As we have m data points in $\mathcal{D}$, we can get m PLRs $\{\mathbf{r}_1, ..., \mathbf{r}_m\}$ where $\mathbf{r}_i = g_{\mathbf{w_j}}(\mathbf{x_i})$. To distinguish between PLRs of different models, we use $R_j := \{g_{\mathbf{w_j}}(\mathbf{x_1}), ..., g_{\mathbf{w_j}}(\mathbf{x_m})\}$ to represent the PLRs by the j-th model.

Next, FLARE applies MMD [12] on R_i and R_j to test whether the two PLR sequences follow the same distribution. We choose MMD but not other two-sample test methods mainly because the number of PLR points $m = 10$ in one sample is much smaller than the dimensionality $o = 128$ of the data. It is difficult to measure the distribution of such a small sample. Traditional parametric two-sample test methods usually have strong assumptions about the parameters of the population distribution from which the sample is drawn and therefore are not applicable. FLARE utilizes MMD to estimate the distance between two PLR sequences since MMD does not require knowing the PLR distribution. Without loss of generality, the unbiased estimate of MMD between the two PLR sequences R_i and R_j is:

$$\mathrm{MMD}(R_i, R_j) = \frac{1}{m(m-1)}\Bigg[\sum_{a\in R_i}\sum_{b\in R_i, b\neq a} k(\mathbf{a}, b) + \sum_{a\in R_j}\sum_{b\in R_j, b\neq a} k(a,b) - 2\sum_{a\in R_i}\sum_{b\in R_j} k(a,b)\Bigg] \tag{4}$$

where $k(\cdot)$ is a Gaussian kernel function. We expect the empirical test statistic $\mathrm{MMD}(R_1, R_2)$ to be small if R_1 and R_2 are from an identical distribution, and large if the distributions are far apart. We use the shortcut $MMD_{ij} = \mathrm{MMD}(R_i, R_j)$ to represent the MMD between the i-th model's PLRs and j-th model's PLRs.

FLARE utilizes the count of nearest neighbors to estimate the trust score of a model update. *PS* selects the top 50% nearest neighbors for each local model based on the MMD scores. The count ct_i for $\mathbf{w_i}$ increases by one once $\mathbf{w_i}$ is selected by any $\mathbf{w_j}(j \neq i)$. The count ct_i value indicates the trustworthiness degree. We then use the softmax function with temperature to transform the count value into a trust score:

$$S_i = \frac{\exp(ct_i/\tau)}{\sum_{k=1}^{n}\exp(ct_k/\tau)}, \tag{5}$$

where τ is the temperature parameter. S_i is in the interval of $[0, 1]$ and $\sum_i^n S_i = 1$. For a sequence of $\{ct_i\}_{i\in[n]}$, a larger τ will output more even trust scores. We can select a smaller τ to highlight benign model updates and reduce the weights of suspicious model updates. We use $\tau = 1$ in our paper.

Algorithm 1 FLARE Algorithm

Input: n local updates $\{\delta_1, \delta_2, ..., \delta_n\}$, m auxiliary data points $\{\mathbf{x}_1, ...\mathbf{x}_m\}$, global model θ, maximum iteration T.
Output: global model θ.

while $t < T$ **do**
  Local models: $\mathbf{w_1}, ..., \mathbf{w_n} \leftarrow \delta_1 + \theta, ..., \delta_n + \theta$.
  for $i < n, i < j < n$ **do**
    $R_i \leftarrow [g_{\mathbf{w}_i}(\mathbf{x_1}), ..., g_{\mathbf{w}_i}(\mathbf{x_m})]$ # i-th model's PLRs.
    $R_j \leftarrow [g_{\mathbf{w}_j}(\mathbf{x_1}), ..., g_{\mathbf{w}_j}(\mathbf{x_m})]$ # j-th model's PLRs
    $MMD_{ij} = MMD_{ji} \leftarrow \text{MMD}(R_i, R_j)$
  end for
  $k = \text{round}(n * 50\%)$ # 50% of the number of updates
  for $i < n$ **do**
    $IDs \leftarrow \text{argsort}(MMD_{i1}, ...MMD_{in})$
    $Neighb_i \leftarrow$ first k elements in IDs
  end for
  # Count times that i is selected as others' neighbors.
  $ct_1, ..., ct_n \leftarrow \text{counting}(Neighb_1, ..., Neighb_n)$
  $\theta \leftarrow \theta + \sum_{i=1}^{n} \frac{e^{ct_i}}{\sum_{k=1}^{n} e^{ct_k}} \delta_i$ # update global model
end while

An alternative scheme is to use the average MMD value of one model to other models to estimate its trust score. The reason why we select the nearest-neighbor-count-based scheme but not the average-MMD-based scheme is shown as follows. The nearest-neighbor-count-based scheme is more resilient to collusive attackers than the average-MMD-based scheme. Colluded attackers can produce nearly the same PLRs, thus resulting in extremely small MMD with each other. In this way, the final average MMD of an attacker may be smaller than benign models, making the detection scheme useless. On the other hand, the counts of the nearest neighbor can deal with this type of collusion when attackers are less than 50% of all clients.

Finally, we aggregate model updates weighted by their trust scores and use it to update the global model by

$$\theta \leftarrow \theta + \sum_{i=1}^{n} S_i \delta_i. \tag{6}$$

where n is the number of local model updates received by *PS*.

6 IMPLEMENTATION AND EXPERIMENTAL SETTINGS

We implement MPAs and FLARE on the TensorFlow platform. We run all the experiments on a server equipped with an Intel Core i7-8700K CPU 3.70GHz×12, a GeForce RTX 2080 Ti GPU, and Ubuntu 18.04.3 LTS. We implement the following four types of MPAs: `Attack-Krum-Untargeted`, `Attack-TM-Untargeted` [10], `Attack-Krum-Backdoor`, and `Attack-Coomed-Backdoor` [4]. We also implement defenses, including `Krum` [5], `Coomed`, `TrimmedMean` [34], `Bulyan` [9], and `FLTrust` [6], as baselines for comparison.

6.1 Experimental Setting

The default number of clients in the studied FL system is $n = 10$, and the ratio of selected clients in each FL iteration is 1.0. The default number of malicious client(s) is one in backdoor attacks and three in untargeted attacks following the setting of MPAs in [4, 10]. We divide the dataset evenly into n subsets and distribute them to clients. The *PS* has an auxiliary dataset containing $m = 10$ clean data points from one class. Each client manages a local model (i.e., VGGNet [28]) and trains the local model using an Adam optimizer with learning rate 0.001. A client trains its local model for five epochs before submitting the model updates. The number of total FL iterations is $T = 20$. The testing accuracy in attack-free model is as shown in Table 2. We run each experiment *three* times and show the average performance. We use three different datasets including fMNIST dataset [32], CIFAR-10 dataset [19] and Kather dataset [17] to evaluate FLARE. The details of the three datasets and corresponding model architectures are depicted in Appendix.

Table 2: Model Accuracy (%) in attack-free scenario.

Dataset	FedAvg	Krum	Coomed	TMean	Bulyan	FLARE
fMNIST	91.77	88.68	91.55	91.61	91.45	91.58
CIFAR-10	69.58	55.190	69.31	69.35	68.56	67.00
Kather	78.83	75.1	76.6	78.33	75.1	78.23

6.2 Evaluation Metrics.

We aim to answer two questions: Is FLARE effective in defending against MPAs by reducing the attack success rate? Can FLARE maintain high accuracy on clean data? Therefore, we show model *confidence* of target label, *attack success rate* (ASR), and model accuracy (Acc) of clean data for evaluating our defense against backdoor attacks. The model confidence in target label c_t denoted by q_t is a commonly used metric for backdoor attacks [4]. ASR is defined as the number of test inputs predicted as the target label over the total number of targeted inputs. Here, a targeted input means an input with a backdoor trigger. The metric for untargeted attacks is different as the attacking goal is different, and model accuracy is the only evaluation metric for untargeted attacks.

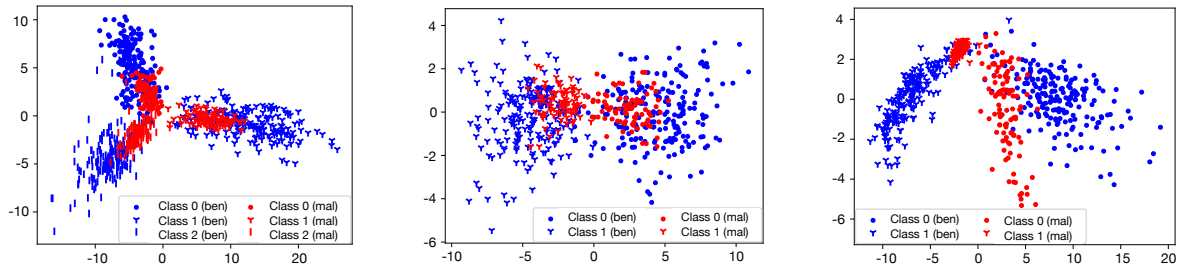
7 EVALUATION RESULTS

7.1 Backdoor Attacks

7.1.1 Attack strategy. In a backdoor attack, the attacker aims to control the predictions on target input without degrading the overall prediction performance on other input data records. As an example of backdoor MPA, several hospitals aim to train a tumor tissue detector through FL. A backdoor attacker injects malicious updates through FL iterations to mislead the FL model to classify tumor tissue as normal tissue. Next, we depict two state-of-the-art backdoor MPAs that are used for evaluation.

`Attack-Krum-Backdoor` [4]: The adversary crafts malicious local models to backdoor FL under *Krum* aggregation rule. This attack is subtle as the malicious parameters are close to those of benign ones and seem innocuous. As a result, the chances are high that the crafted malicious model parameters are accepted by *Krum*. The objective function is:

$$\arg\min_{\delta_{mal}} \quad L(\mathcal{D}_{mal}) + \lambda L(\mathcal{D}_{train}) + \rho \|\delta_{mal} - \bar{\delta}_{ben}\|, \tag{7}$$



(a) fMNIST (b) CIFAR-10 (c) Kather

Figure 6: PLRs of different classes without attack (blue) or under backdoor attack (red).

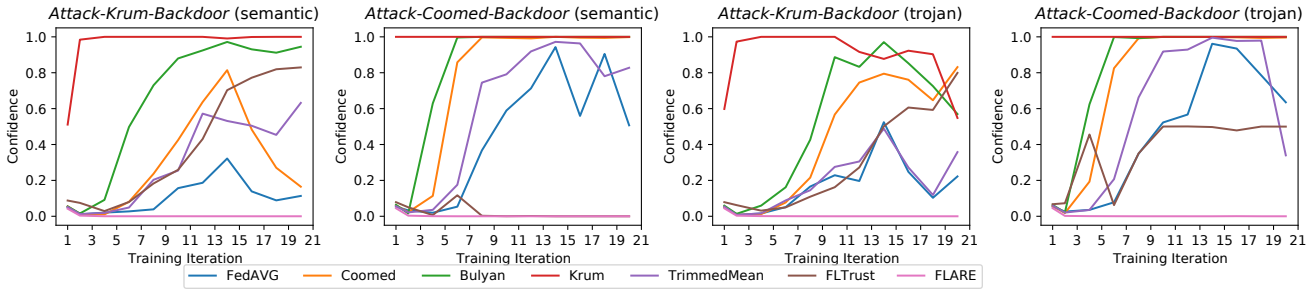


Figure 7: Model Confidence in targeted inputs under backdoor MPAs (fMNIST dataset). The first two are semantic backdoor attacks, and the last two are trojan backdoor attacks.

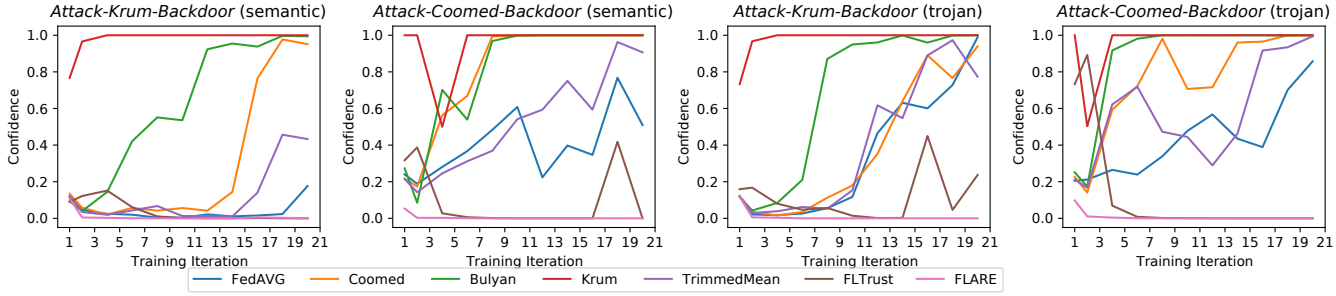


Figure 8: Model confidence on targeted label under backdoor MPAs on CIFAR-10 dataset.

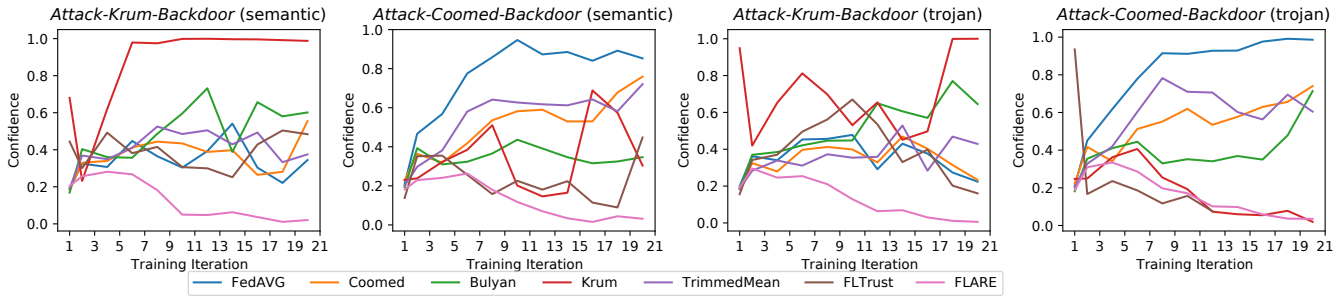
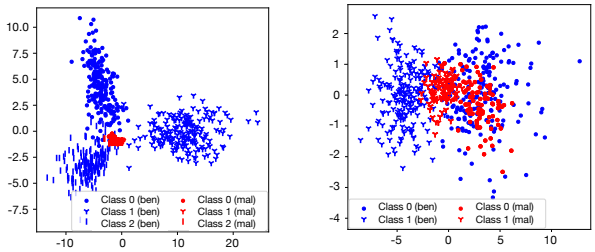


Figure 9: Model confidence on targeted label under backdoor MPAs on Kather dataset.

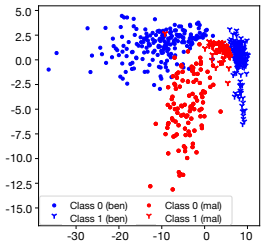
Table 3: Model Accuracy (%) on clean data and ASR (%) on *targeted* data.

Attack Name	Dataset	FedAvg		Krum		Coomed		TrimmedMean		Bulyan		FLTrust		FLARE	
		Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR
Attack-Krum	fMNIST	91.6	8.3	87.9	98.3	91.6	45.0	91.5	43.3	91.4	71.6	**91.7**	40.0	91.2	**0**
-Backdoor	CIFAR-10	**68.7**	26.7	48.1	100	67.8	53.3	68.4	33.3	67.2	65.0	67.0	**0**	66.4	**0**
(semantic)	Kather	**79.9**	41.6	50.9	86.7	79.3	33.3	79.4	41.7	79.5	56.7	50.5	19.4	76.7	**0**
Attack-Coomed	fMNIST	91.5	58.3	88.0	98.3	**91.7**	78.3	91.6	68.3	91.5	85.0	91.5	**0**	91.4	**0**
-Backdoor	CIFAR-10	**68.0**	56.7	50.1	96.7	67.8	85.0	67.7	53.3	66.1	88.3	66.3	20.3	65.2	**0**
(semantic)	Kather	75.3	91.7	64.8	51.6	**78.6**	78.3	78.2	75.0	78.5	46.7	76.2	7.2	77.8	**0**



(a) fMNIST

(b) CIFAR-10



(c) Kather

Figure 10: The PLRs of 3 classes without attack (blue) or under untargeted attack (red) in fMNIST dataset, CIFAR dataset, and Kather dataset, respectively.

where the main goal is to minimize the loss on backdoor inputs, denoted by $L(\mathcal{D}_{mal})$. Meanwhile, the attack aims to minimize the loss $L(\mathcal{D}_{train})$ to improve the accuracy of clean samples. Additionally, the attack also minimizes the distance $\|\delta_{mal} - \bar{\delta}_{ben}\|$ between the malicious update and average benign updates to be stealthy. By solving this objective, this attack can mislead the global model to output target labels for target inputs while keeping stealthy.

`Attack-Coomed-Backdoor` [4]: The objective function of this attack is the same as that of `Attack-Krum-Backdoor` (i.e., Eq. (7)). To defeat *coordinate median* aggregation rule, the local training process at attackers is a little different from `Attack-Krum-Backdoor` (see [4] for details).

According to the need of physically injecting a trigger, backdoor attacks are categorized into two types, i.e., trojan backdoor attack and semantic backdoor attack. In the trojan attack, attackers need to physically inject a backdoor/trigger to an ML model by modifying all or a subset of the training data. Different from the trojan attack, *semantic* backdoor attack involves no physically injected trigger in inputs. In the *semantic* backdoor attack, the trigger is a semantic feature included in an original image. For instance, the 'stripes' can be a semantic trigger for the apparel classification problem. The attacker attaches the label 'sweater' to images containing 'stripes'. As a result, any apparel with 'stripes' will be classified as sweaters. We extend the two backdoor MPAs (i.e., `Attack-Krum-Backdoor` and `Attack-Coomed-Backdoor`) into four attacks including `Attack-Krum-Backdoor(semantic)/(trojan)` and `Attack-Coomed-Backdoor(semantic)/(trojan)`.

7.1.2 FLARE Performance. Table 3 shows the performance of FLARE against two semantic backdoor MPAs. The name of an attack on the left column, such as `Attack-Krum-Backdoor`, contains the targeting BRAR (i.e., `Krum`) to attack. From Table 3 we can see that an MPA can successfully attack not only its targeting BRAR but also undermine other BRARs. On the contrary, our proposed FLARE can reduce the ASR to a small value close to zero across various datasets. FLARE outperforms BRARs and FLTrust by achieving a lower ASR. As for accuracy on clean data, FLARE obtains a slightly lower accuracy than other baselines. This is because a malicious update contains both the poisoning knowledge and useful knowledge from its own clean data. In FLARE, *PS* assigns small weights to the updates from malicious clients thus it achieves a slightly lower accuracy on clean data.

We plot the global model's confidence on targeted inputs in Figure 7. The first two subfigures of Figure 7 show the model confidence under two semantic backdoor attacks and the last two for trojan backdoor attacks. Under `Attack-Krum-Backdoor (semantic)`, we can see that the model confidence of `Krum` rises to 1.0 very fast, indicating that the attack succeeds at the early stage of FL. Under the same attack, the model confidence of other BRARs such as `Bulyan`, `Coomed`, `Trimmedmean` also increases along with the learning process, indicating that this attack succeeds in attacking all the BRARs. We can see that the confidence fluctuates along the training process because of the randomness in the distributed learning and the aggregation rules. Note that all the results shown are the average value of three runs. On the contrary, FLARE results in very steady and small confidence scores for the targeted input under MPAs, meaning that FLARE successfully defends against the four backdoor attacks. We achieve similar results on the CIFAR-10 dataset and Kather dataset as shown in Figure 8 and Figure 9 respectively. For both two datasets, FLARE achieves the lowest confidence in the final (i.e., 20th) iteration. We can see the model confidence is close to 0 at the final iteration, which indicates that FLARE successfully defenses against these attacks.

Table 4: Model Accuracy (%) under *untargeted* attacks.

Attack Name	Dataset	FedAvg	Krum	Coomed	TrimmedMean	Bulyan	FLTrust	FLARE
Attack-Krum -Untargeted	fMNIST	59.2	6.93	79.8	89.7	86.2	**91.2**	90.9
	CIFAR-10	62.4	10.4	61.6	62.8	62.5	66.4	**66.5**
	Kather	69.0	12.34	66.1	66.4	69.8	11.6	**76.4**
Attack-TrimmedMean -Untargeted	fMNIST	87.1	87.9	80.3	61.1	89.5	90.7	**91.1**
	CIFAR-10	58.8	53.4	60.3	49.2	64.0	64.4	**67.5**
	Kather	18.0	70.8	64.2	22.3	74.1	10.9	**77.5**

7.2 Untargeted Attacks

7.2.1 Attack Strategy. In untargeted attacks, the attacker aims to degrade model performance by preventing the global model from convergence or leading the global model to converge to a local optimum that yields a high testing error rate. Next, we summarize two state-of-the-art untargeted MPAs that are used for evaluation.

`Attack-Krum-Untargeted` [10]: The adversary aims to craft k ($k \geq 1$) malicious local models to attack *Krum*. To achieve this, the attacker applies "directed deviation" to the global model parameters, which moves the parameters along the opposite direction of the attack-free one. The attack is formalized as:

$$\begin{aligned} \max \quad & \lambda \\ \text{s.t.} \quad & \mathbf{w}'_1 = Krum(\mathbf{w}'_1, ..., \mathbf{w}'_k, \mathbf{w}_{(k+1)}, ..., \mathbf{w}_n), \\ & \mathbf{w}'_1 = \theta - \lambda \mathbf{s}, \\ & \mathbf{w}'_i = \mathbf{w}'_1, \text{for } i = 2, 3, ..., k. \end{aligned} \tag{8}$$

where $\mathbf{w}'_i$ ($\forall i \in [k]$) represents the weights of a poisoned model; $\mathbf{w}_i$ ($\forall i \in \{k+1, ..., n\}$) represents the weights of a benign model; θ represents the current global model; and $\mathbf{s}$ represents the sign of the average weights of all benign models. The directed deviation between the crafted model $\mathbf{w}'_1$ and global model θ is $\lambda \mathbf{s}$. The objective of this attack is to maximize λ in order to increase the error rate of FL.

`Attack-TM-Untargeted` [10]: Similar to `Attack-Krum-Untargeted`, the high-level idea is to deviate the global model toward the opposite direction of the attack-free model. Assume that the benign weight in the j-th coordinate is in the range $[w_{j,min}, w_{j,max}]$. To attack *TrimmedMean*, the j-th coordinate of the crafted model should be in the same range (i.e., $[w_{j,min}, w_{j,max}]$) of attack-free model. In this attack, the attacker develops the following heuristic algorithm. In specific, the j-th coordinate is crafted by sampling a value around $w_{j,max}$ if the sign of the average weight is negative (i.e., $s_j = -1$). On the contrary, the j-th coordinate is generated by sampling around $w_{j,min}$ if $s_j = 1$. As a result, the attack can flip the sign of some coordinates of the average weights. Similar to `Attack-Krum-Untargeted`, this attack aims to increase the testing error rate of the global model.

7.2.2 FLARE Performance. We show the testing accuracy of the final global model under untargeted MPAs in Table 4. We can see that the untargeted attacks can not only successfully spoil the target BRAR, but also make other BRARs less effective. FLARE successfully defends against untargeted MPAs by achieving testing accuracy much higher than other baselines. However, the accuracy of FLARE is still lower than the attack-free scenario (see Table 2). The possible reason is the same as backdoor MPAs, i.e., a lack of helpful knowledge from malicious updates. FLTrust achieves comparable accuracy with FLARE in the fMNIST dataset and CIFAR-10 dataset, but it achieves a very low accuracy in the Kather dataset. A possible reason is that FLTrust assigns a trust score to a local model update based on its cosine similarity with the benign model trained at *PS*. The cosine similarity is meaningless when the dimension of the model parameters is huge.

7.3 Performance in Various FL settings

In order to demonstrate the robustness of FLARE against MPAs, we evaluate FLARE in various settings. Let's take FLARE's performance against `attack-Krum-backdoor` as an example.

7.3.1 Client Number, Attacker Number and Auxiliary Data Points. In Figure 11(a), we vary the number of clients while keeping the malicious clients as 10% of the total clients. The ASR without defense stays higher than 0.8. The ASR after applying FLARE is close to zero regardless of the number of clients. The results indicate that FLARE effectively defended against MPAs in FL systems with different numbers of clients. In Figure 11(b), the percentage p of malicious clients is from 5 to 30. We can see the ASR under FLARE remains close to 0 when $p < 30\%$, implying that FLARE is highly effective when the malicious clients are fewer than 30% of the total clients. It is challenging to detect malicious clients when they are more than 30% of the total clients. We further examine the impact of the size m of the $\mathcal{D}$ on FLARE and find that FLARE is effective when $m \geq 7$.

7.3.2 Non-I.I.D. Data. We follow the setting in [25] to generate the non-i.i.d. datasets among distributed clients. we assign every client samples from exactly $c_{sel} < c$ classes of the data set where c is the total categories. The data splits are nonoverlapping and balanced, such that every client ends up with the same number of data points. We show the ASR of the two backdoor attacks in Table 5. FLARE obtains the smallest ASR for the two attacks, demonstrating its resilience against MPAs in non-i.i.d. scenario.

Table 5: ASR in non-i.i.d. scenario in fMNIST.

Attack	Krum	Coomed	TMean	Bulyan	FLARE
Attack-Krum-Backdoor	0.750	0.533	0.133	0.716	**0.016**
Attack-Coomed-Backdoor	1.00	0.867	0.750	0.883	**0.050**

7.4 Defending against Adaptive Attack

In a more challenging scenario, an attacker can adaptively alter their attack methods to defeat the defense with the knowledge of

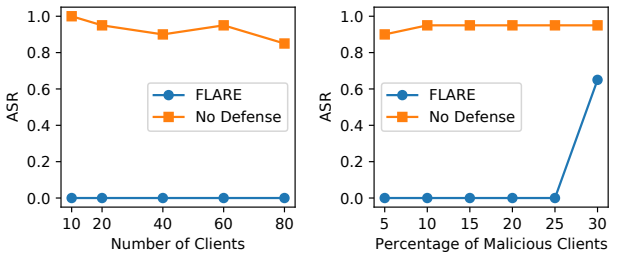


Figure 11: ASR against `Attack-Krum-Backdoor` in fMNIST dataset. (a) vary the number of clients when fixing the percentage of malicious clients as 0.1; (b) vary the percentage malicious clients when fixing total clients as 20.

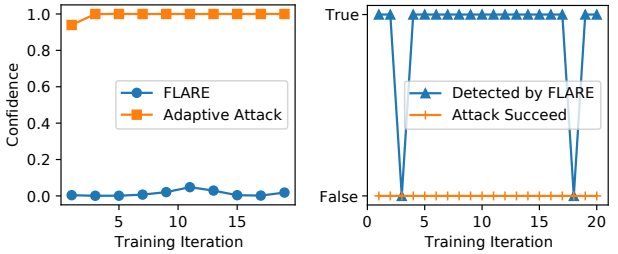


Figure 12: FLARE Performance against adaptive attack.

the defense strategy. Under this attack, we assume the attacker knows the defense strategy of FLARE.

The adaptive attack follows a strategy explained below. In order to bypass FLARE, the attacker crafts its model to produce PLRs similar to PLRs of benign local models. The attack objective is shown as:

$$\arg\min_{\delta_{mal}} \quad L(\mathcal{D}_{mal}) + \lambda L(\mathcal{D}_{train}) + \rho\|\delta_{mal} - \bar{\delta}_{ben}\| + \eta d_{plr}. \tag{9}$$

where $L(\mathcal{D}_{mal})$ is the loss on targeted inputs, $L(\mathcal{D}_{train})$ is the loss on the clean data, and $\|\delta_{mal} - \bar{\delta}_{ben}\|$ is the distance between malicious model updates and average benign model updates, and d_{plr} represents the distance between the PLRs of malicious model and the average PLRs of benign models. This formula originates from [4]. This attack misleads the global model to output target labels for chosen inputs while hiding its maliciousness. The original attack uses components $L(\mathcal{D}_{train})$ and $\|\delta_{mal} - \bar{\delta}_{ben}\|$ to achieve the stealthiness. Compared to the original attack in [4], we add one more component d_{plr} to the objective function which can be represented as:

$$d_{plr} = \sum_{i \in [c]} (\overline{PLR}(\delta_{mal}, \mathcal{D}^i_{train}) - \overline{PLR}(\bar{\delta}_{ben}, \mathcal{D}^i_{train})), \tag{10}$$

where c is the total number of classes of the classifier, and D^i_{train} is the training dataset of class i. For each class i, $\overline{PLR}(\delta_{mal}, \mathcal{D}^i_{train})$ denotes the average value of PLRs of the malicious model on the training data D^i_{train}, and $\overline{PLR}(\bar{\delta}_{ben}, \mathcal{D}^i_{train}))$ denotes the average value of PLRs of benign model $\bar{\delta}_{ben}$ on the same dataset D^i_{train}.

Figure 12(a) shows the performance of FLARE against the adaptive attack. When no defense is applied, the adaptive attack itself achieves a high attack success rate by achieving confidence larger than 0.9. After FLARE is deployed into the learning process, the confidence remains less than 0.1, and the ASR decreases to 0. Such results demonstrate the effectiveness of FLARE for defending against the adaptive attack. In Figure 12(b), we show two flags: one indicates whether the attack is detected, the other indicates whether the attack makes a misclassification. Note that here 'detected' means that a malicious model update obtains a trust score lower than average. We can see that FLARE fails to detect malicious clients in Iteration 3 and 18. Meanwhile, the adaptive attack does not succeed in these two rounds either, which means that the maliciously crafted model in Iteration 3 or 18 becomes innocuous. Such an observation confirms that it is difficult for adaptive attacks to evade FLARE and achieve malicious goals simultaneously.

We also analyze the computation complexity of FLARE. We find that the computation overhead of FLARE depends on PLR calculation and MMD execution. FLARE's time complexity can be represented as $O(n \cdot t_{plr} + n^2 \cdot t_{mmd})$ where n is the number of selected clients in each iteration, t_{plr} is PLR computation time per client, and t_{mmd} is the computation time for one MMD test. In our case, t_{plr} is small as there are only 10 MMD test samples. The experimental execution time of FLARE is 3.2 seconds when $n = 10$ and 28.3 seconds when $n = 50$, indicating that FLARE is relatively computation-efficient.

8 CONCLUSIONS

In this paper, we propose a robust aggregation algorithm FLARE to protect FL against MPAs. Through analysis and experimental visualization, we demonstrate that the PLR vector has high potentials in differentiating malicious/poisonous models from the benign ones. Based on the PLR technique, FLARE effectively minimizes the impact of malicious/poisonous models on the final aggregation by assigning low trust scores to those with diverging PLRs. Through a comprehensive evaluation, we show that FLARE significantly outperforms existing defenses (i.e., BRARs and FLTrust) in defending against state-of-the-art MPAs, including semantic backdoor attacks, trojan attacks, and untargeted attacks on three popular datasets. Furthermore, FLARE also shows its effectiveness amid non-i.i.d. data and adaptive attacks, demonstrating the applicability to challenging real-world scenarios.

ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research under grant N00014-19-1-2621, the US National Science Foundation under grants CNS-1837519 and CNS-19169026, the Army Research Office under grant W911NF-20-1-0141, and the Virginia Commonwealth Cyber Initiative (CCI).

REFERENCES

[1] Sebastien Andreina, Giorgia Azzurra Marson, Helen Möllering, and Ghassan Karame. 2021. Baffle: Backdoor detection via feedback-based federated learning. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 852–863.

[2] Sana Awan, Bo Luo, and Fengjun Li. 2021. Contra: Defending against poisoning attacks in federated learning. In *European Symposium on Research in Computer Security*. Springer, 455–475.

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

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

[5] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 118–128.

[6] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. 2021. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. *Network and Distributed Systems Security Symposium NDSS* (2021).

[7] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2021. Provably Secure Federated Learning against Malicious Clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6885–6893.

[8] Yudong Chen, Lili Su, and Jiaming Xu. 2017. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1, 2 (2017), 1–25.

[9] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Louis Alexandre Rouault. 2018. The Hidden Vulnerability of Distributed Learning in Byzantium. In *International Conference on Machine Learning*.

[10] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. 2020. Local model poisoning attacks to Byzantine-robust federated learning. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 1605–1622.

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

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[13] Rachid Guerraoui, Sébastien Rouault, et al. 2018. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*. PMLR, 3521–3530.

[14] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).

[15] Peter J Huber. 2004. *Robust statistics*. Vol. 523. John Wiley & Sons.

[16] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning* 14, 1 (2021). https://doi.org/10.1561/2200000083

[17] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. 2016. Multi-class texture analysis in colorectal cancer histology. *Scientific reports* 6 (2016), 27988.

[18] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[19] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report. Citeseer.

[20] Suyi Li, Yong Cheng, Wei Wang, Yang Liu, and Tianjian Chen. 2020. Learning to Detect Malicious Clients for Robust Federated Learning. *arXiv preprint arXiv:2002.00211* (2020).

[21] Yunlong Mao, Xinyu Yuan, Xinyang Zhao, and Sheng Zhong. 2021. Romoa: Robust model aggregation for the resistance of federated learning to model poisoning attacks. In *European Symposium on Research in Computer Security*. Springer, 476–496.

[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics (AISTATS 17)*. 1273–1282.

[23] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help?. In *Advances in Neural Information Processing Systems*. 4694–4703.

[24] Thien Duc Nguyen, Phillip Rieger, Markus Miettinen, and Ahmad-Reza Sadeghi. 2020. Poisoning attacks on federated learning-based IoT intrusion detection system. In *Proc. Workshop Decentralized IoT Syst. Secur.(DISS)*. 1–7.

[25] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. 2019. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems* 31, 9 (2019), 3400–3413.

[26] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. 2019. Regularization via mass transportation. *Journal of Machine Learning Research* 20, 103 (2019), 1–68.

[27] Shiqi Shen, Shruti Tople, and Prateek Saxena. 2016. Auror: Defending against poisoning attacks in collaborative deep learning systems. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*. 508–519.

[28] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[29] Jinhyun So, Başak Güler, and A Salman Avestimehr. 2020. Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications* (2020).

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

[31] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. 2019. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications* 37, 6 (2019), 1205–1221.

[32] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[33] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. 2020. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*.

[34] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*. PMLR, 5650–5659.

[35] Ying Zhao, Junjun Chen, Jiale Zhang, Di Wu, Jian Teng, and Shui Yu. 2019. PDGAN: A novel poisoning defense method in federated learning using generative adversarial network. In *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 595–609.

A PROOF OF PROPOSITION 1

Proposition 1. In a c-class NN classifier where the last two layers are fully connected and the last layer is a softmax layer, the output probability of any two class k and l ($1 \leq k, l \leq c$) satisfy that $q_k > q_l$ if

$$||\mathbf{r} - \omega_l||_2 - ||\mathbf{r} - \omega_k||_2 \geq C_{kl} \tag{11}$$

where $\mathbf{r}$ represents the penultimate layer value and ω_k is the weights connecting to the k-th neuron of the output layer. $||\mathbf{r} - \omega_k||_2$ denotes the Euclidean distance between $\mathbf{r}$ and template ω_k, i.e., $||\mathbf{r} - \omega_k||_2 = \mathbf{r}^T\mathbf{r} - 2\mathbf{r}^\mathbf{T}\omega_k + \omega_k^T\omega_k$. C_{kl} is a constant and $C_{kl} = \omega_l^T\omega_l - \omega_k^T\omega_k$.

PROOF. The prediction probability vector of a c-class classifier on an input is denoted by $[q_1, q_2, ..., q_c]$ where $\sum_c q_k = 1$. The probability q_k can be represented by

$$q_k = \frac{\exp(\mathbf{r}^T\omega_k)}{\sum_{i=1}^{c}\exp(\mathbf{r}^T\omega_i)}. \tag{12}$$

where $\mathbf{r}$ represents the penultimate layer value and ω_k is the weights connecting to the k-th neuron of the output (i.e., softmax) layer. The Euclidean distance between the penultimate layer representation $\mathbf{r}$ and template ω_k is

$$||\mathbf{r} - \omega_k||_2 = \mathbf{r}^T\mathbf{r} - 2\mathbf{r}^\mathbf{T}\omega_k + \omega_k^T\omega_k. \tag{13}$$

We can derive the logit $\mathbf{r}^\mathbf{T}\omega_k$ from Eq. (13) by

$$\mathbf{r}^\mathbf{T}\omega_k = -\frac{1}{2}||\mathbf{r} - \omega_k||_2 + \frac{1}{2}\mathbf{r}^T\mathbf{r} + \frac{1}{2}\omega_k^T\omega_k. \tag{14}$$

We rewrite q_k by substituting Eq. (14) for the logit $\mathbf{r}^{\mathbf{T}}\omega_k$ in Eq. (12)

$$\begin{aligned} q_k &= \frac{\exp(-\frac{1}{2}||\mathbf{r}-\omega_k||_2)\cdot\exp(\frac{1}{2}\mathbf{r}^T\mathbf{r})\cdot\exp(\frac{1}{2}\omega_k^T\omega_k)}{\sum_i^c \exp(-\frac{1}{2}||\mathbf{r}-\omega_i||_2)\cdot\exp(\frac{1}{2}\mathbf{r}^T\mathbf{r})\cdot\exp(\frac{1}{2}\omega_i^T\omega_i)} \\ &= \frac{\exp(-\frac{1}{2}||\mathbf{r}-\omega_k||_2)\cdot\exp(\frac{1}{2}\omega_k^T\omega_k)}{\sum_i^c \exp(-\frac{1}{2}||\mathbf{r}-\omega_i||_2)\cdot\exp(\frac{1}{2}\omega_i^T\omega_i)} \\ &= \frac{\exp(-\frac{1}{2}||\mathbf{r}-\omega_k||_2+\frac{1}{2}\omega_k^T\omega_k)}{\sum_i^c \exp(-\frac{1}{2}||\mathbf{r}-\omega_i||_2+\frac{1}{2}\omega_i^T\omega_i)} \end{aligned} \tag{15}$$

We use q_k and q_l to represent the probability of the k-th and l-th class. If $q_k \geq q_l$, the inequality (16) must be satisfied.

$$-\frac{1}{2}||\mathbf{r}-\omega_k||_2+\frac{1}{2}\omega_k^T\omega_k \geq -\frac{1}{2}||\mathbf{r}-\omega_l||_2+\frac{1}{2}\omega_l^T\omega_l \tag{16}$$

This inequality can be rewritten as

$$||\mathbf{r}-\omega_l||_2-||\mathbf{r}-\omega_k||_2 \geq \omega_l^T\omega_l-\omega_k^T\omega_k \tag{17}$$

The right hand is a constant, and we can use C_{kl} to represent this constant, i.e., $C_{kl}=\omega_l^T\omega_l-\omega_k^T\omega_k$. The Eq. (17) is rewritten as

$$||\mathbf{r}-\omega_l||_2-||\mathbf{r}-\omega_k||_2 \geq C_{kl} \tag{18}$$

Eq. (18) indicates that if the distance between representation $\mathbf{r}$ and template l is larger than the distance between representation $\mathbf{r}$ and template k by a constant C_{kl}, then the probability of assigning the input as class k is larger than class l.

B PROOF OF PROPOSITION 2

Proposition 2. The mapping function $\sigma : \mathbf{r} \in \mathbb{R}^o \rightarrow \mathbf{q} \in \mathbb{R}^c$ maps a PLR to a probability vector as discussed above. For any two PLRs $\mathbf{r}_1, \mathbf{r}_2$, we have

$$\|\mathbf{q}^1-\mathbf{q}^2\|_2 \leq \|\Omega\|_2\|\mathbf{r}_1-\mathbf{r}_2\|_2, \tag{19}$$

where $\|.\|_2$ is the L2 norm operator, $\mathbf{r}_1$ and $\mathbf{r}_2$ are the PLR of two input x_1 and x_2 respectively. $\mathbf{q}^1$ and $\mathbf{q}^2$ are the output probability vector for input x_1 and x_2 respectively.

PROOF. Here we use $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^c$ to represent the logits calculated from $\mathbf{r}_1, \mathbf{r}_2 \in \mathbb{R}^o$, i.e., $\mathbf{z}_1 = \Omega^T r_1$ and $\mathbf{z}_2 = \Omega^T \mathbf{r}_2$, where $\Omega \in \mathbb{R}^{o\times c}$ denotes the weights from the penultimate layer to the last layer.

Here we use $\zeta : \mathbb{R}^c \rightarrow \mathbb{R}^c$ to represent the softmax function, then we have the probability vector $\mathbf{q}^1 = \zeta(\mathbf{z}_1)$ and $\mathbf{q}^2 = \zeta(\mathbf{z}_2)$. Based on that the Lipchitz modulus of softmax function ζ is less than one [26], thus we have

$$\|\mathbf{q}^1-\mathbf{q}^2\|_2 \leq \|\mathbf{z}_1-\mathbf{z}_2\|_2. \tag{20}$$

By substituting $\mathbf{z}_1$ and $\mathbf{z}_2$ with $(r_2)^T\Omega$ and $(r_1)^T\Omega$ respectively, we have

$$\|\mathbf{q}^1-\mathbf{q}^2\|_2 \leq \|(\mathbf{r}_1)^T\Omega-(\mathbf{r}_2)^T\Omega\|_2 \leq \|\Omega\|_2\|\mathbf{r}_1-\mathbf{r}_2\|_2 \tag{21}$$

C NEURAL NETWORK ARCHITECTURE

The detail of the three datasets is shown in the following, and the corresponding neural networks are shown in Table 6, Table 7, and Table 8. Note that we resize the images in the Kather dataset from $150\times150\times3$ to $128\times128\times3$ before feeding them into the VGGNet. The batch size used in the fMNIST dataset is 64 and the batch size used in the CIFAR-10 dataset and the Kather dataset is 32.

fMNIST consists of a training set of 60,000 records and a test set of 10,000 records. Each data record is a 28×28 grayscale image, associated with a label from 10 classes, including T-shirt, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot.

CIFAR-10 consists of 60,000 32x32 colour images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images. Each image is from one of the ten classes, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Kather is a collection of textures in colorectal cancer histology. It consists of 5,000 records and each one is a 150×150 histological image. Each image belongs to one of the eight tissue categories, including Tumor, Stroma, Complex, Lympho, Debris, Mucosa, Adipose, and Empty.

Table 6: Neural network architecture for fMNIST dataset.

	Input	Filter	Stride	Output	Activation
conv2d_1	28×28×1	64×5×5	1	24×24×64	ReLU
conv2d_2	24×24×64	64×5×5	1	20×20×64	ReLU
dropout_1	20×20×64	Dropout, 0.25	/	20×20×64	/
flatten_1	20×20×64	/	/	25600	/
dense_1	25600	/	/	128	Relu/
dropout_2	128	Dropout, 0.5	/	128	/
dense_2	128	/	/	10	/

Table 7: Neural network architecture for CIFAR-10 dataset.

	Input	Filter	Stride	Output	Activation
conv2d_1	32×32×3	64×3×3	1	30×30×64	ReLU
max_pooling2d_1	30×30×64	MaxPooling2D, 2×2	2	15×15×64	/
conv2d_2	15×15×64	64×3×3	1	13×13×64	ReLU
max_pooling2d_2	13×13×64	MaxPooling2D, 2×2	2	6×6×64	/
conv2d_3	6×6×64	64×3×3	1	4×4×64	ReLU
max_pooling2d_3	4×4×64	MaxPooling2D, 2×2	2	2×2×64	/
flatten_1	2×2×64	/	/	256	/
dense_1	256	/	/	128	/
dense_2	128	/	/	10	/

Table 8: Neural network architecture for Kather dataset.

	Input	Filter	Stride	Output	Activation
conv2d_1	128×128×3	64×3×3	1	128×128×64	ReLU
max_pooling2d_1	128×128×64	MaxPooling2D, 2×2	2	64×64×64	/
conv2d_2	64×64×64	64×3×3	1	64×64×64	ReLU
max_pooling2d_2	64×64×64	MaxPooling2D, 2×2	2	32×32×64	/
conv2d_3	32×32×64	64×3×3	1	32×32×64	ReLU
max_pooling2d_3	32×32×64	MaxPooling2D, 2×2	2	16×16×64	/
conv2d_4	16×16×64	64×3×3	1	16×16×64	ReLU
max_pooling2d_4	16×16×64	MaxPooling2D, 2×2	2	8×8×64	/
conv2d_5	8×8×128	64×3×3	1	8×8×128	ReLU
max_pooling2d_5	8×8×128	MaxPooling2D, 2×2	2	4×4×128	/
flatten_1	4×4×128	/	/	2048	/
dense_1	2048	/	/	128	/
dense_1	128	/	/	8	/