Federated Resistance Against Adversarial Attacks in Resource-constrained IoT

Mahmoud A. Zaher[1,*], Heba H. Aly[2]

[1] Faculty of Artificial Intelligence, Egyptian Russian University (ERU), Cairo, Egypt
[2] Faculty of computers and information systems, Beni Sief University , Cairo, Egypt
Emails: Mahmoud.zaher@eru.edu.eg; Heba.h.ali@fcis.bsu.edu.eg

Abstract
Federated learning (FL) is a recently evolved distributed learning paradigm that gains increased research attention. To alleviate privacy concerns, FL fundamentally suggests that many entities can cooperatively train the machine/deep learning model by exchanging the learning parameters instead of raw data. Nevertheless, FL still exhibits inherent privacy problems caused by exposing the users' data based on the training gradients. Besides, the unnoticeable adjustments on inputs done by adversarial attacks pose a critical security threat leading to damaging consequences on FL. To tackle this problem, this study proposes an innovative Federated Deep Resistance (FDR) framework, to provide collaborative resistance against adversarial attacks from various sources in a Fog-assisted IIoT environment. The FDR is designed to enable fog nodes to cooperate to train the FDL model in a way that ensures that contributors have no access to the data of each other, where class probabilities are protected utilizing a private identifier generated for each class. The FDR mainly emphasizes convolutional networks for image recognition from the Food-101 and CIFAR-100 datasets. The empirical results have revealed that FDR outperformed the state-of-the-art adversarial attacks resistance approaches with 5% of accuracy improvements.

Keywords: Adversarial Attacks; Federated Learning; Fog Computing; Industrial Internet of Things (IIoT)

I. INTRODUCTION

As a result of the confluence of artificial intelligence (AI) with the Industrial Internet of things (IIoT), industrial applications have undergone a significant transformation, becoming significantly more intelligent and efficient [1]. Deep learning is a subfield of artificial intelligence that has recently been getting more attention from researchers in order to develop data-driven solutions for the IIoT. This is due to deep learning's capacity to learn from, model, and find patterns in IIoT data [2]. Large amounts of data are generated and stored in a wide range of devices as a consequence of the rapid emergence of smart IIoT applications and services [23]. This creates favourable conditions for the training of enhanced and efficient deep learning solutions, such as autonomous driving, navigation systems, home automation, remote surgery, smart farming, intelligent buildings, and so on. The proliferation of deep learning solutions, on the other hand, makes them more susceptible to adversarial attacks, whether they are carried out intentionally or accidentally. The goal of these assaults is to change the model's inputs in a way that will cause it to exhibit behaviours that are incorrect or unpredictable, which could lead to catastrophic results.

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

The vast majority of the existing adversarial attacks focus on misleading deep learning models by providing IIoT inputs that are based on perturbations that are referred to as "adversarial samples." The primary goal of these adversarial examples is to trick the model into producing incorrect outputs, even if these outputs are simple enough for a human to compute on their own. In this context, a number of different strategies for resistance have been proposed, including adversarial training, ensemble diversity, and PuVAE [3, 4]. Despite this, almost none of them are suitable for the context of IIoT due to the dynamic, broad-scale, and diverse character of the environment.

In early time, the deep learning solution was trained on the cloud layer of the IIoT system. This was done so that the training data could be centralizedly taught and was where users' data was gathered. However, in order to be useful, this tactic must first overcome two significant obstacles: the breach of privacy it entails and the significant delay it causes. As a collaborative machine learning paradigm, federated learning (FL) emerged as a solution to these limitations. FL enables training on the neighbourhood of client systems without the need to upload the data to a centralised cloud, which allows for the data's privacy to be protected [5–7]. [FL] was developed to address these limitations. [FL] [Federated Learning] Putting it another way, federated learning gives data owners the ability to train a deep learning solution using their own local data, only share parameters with a third-party trust parity, and deduce categorization outputs without revealing any confidential or sensitive information pertaining to their customers. For the purpose of training federated learning solutions, edge computing has received a lot of investigation. However, edge devices are still susceptible to difficulties caused by a lack of resources, and as a result, lightweight solutions are required. The robust cloud resources can be brought closer to the edge of IIoT networks through the use of fog computing, which can give a solution to this problem [8].

The integration of fog computing and IIoT is turning out to be a promising computing paradigm that enables broad-scale smart IIoT services and applications. This is because fog assisted IIoT systems can expand the computational facilities of IIoT devices by offloading their limited computation tasks to an intermediate and close fog server. [12] Since fog assisted IIoT systems can expand the computational facilities of IIoT devices, they can offload their limited computation tasks to an intermediate and close fog server. In spite of this, regardless of whether cloud servers are used for training or inference, IIoT devices are required to communicate original data to cloud servers, which brings up a number of challenges that cannot be ignored, including network latency and data privacy concerns. In addition, the toughness of an application may be readily breached when a number of distinct IIoT devices all employ the same particular deep learning model. This is because there may be variances in the category of adversarial assaults that are being used. It is therefore a very difficult task to scale up an effective deep learning model for a wide variety of IoT devices while also ensuring that the models are robust to adversarial assaults and that they do not compromise users' privacy.

The following three key issues are addressed in this study. 1) malicious attacks have a cataclysmic effect on the quality of the applications and services provided by industrial organisations. The vast majority of resistance mechanisms centre their attention on a specific class of threats, which makes them inappropriate for widespread implementation in the field of IIoT. The surroundings of IIoT are often dispersed across a variety of geographic regions, and thus are susceptible to a wide variety of hostile attacks. Under these circumstances, IoT devices of the same sort will need to be outfitted with distinct models in order to function well in a variety of environments. When many new adversarial attacks emerge, things get even more difficult for IIoT service providers, as it is tough for them to quickly create a new solution to withstand these attacks. This makes the situation even more precarious. 2) cloud-based services have a well-deserved reputation for being riddled with data privacy flaws. The phenomenal success of federated learning has inspired the FDR to investigate the feasibility of federated training as an alternative to shifting data to the cloud. However, the absence of a trust mechanism often makes the user's data susceptible to be illegally accessed by another user. This is one of the most significant barriers to the widespread approval of federated IIoT applications [9], as it is one of the primary reasons why users are hesitant to use such applications. 3) the bulk of applications for the IIoT are required to keep a high level of service quality. Because of this, it is customarily necessary for the deep learning model to keep its efficiency even as it is being subjected to some adversarial attacks [10]. The correctness of the results, in addition to the amount of time it takes for a response, is the efficiency in this case [7]. The achievements of fog computing served as inspiration for the FDR's decision to train the deep learning model on fog nodes. This shift was made in order to move the computation closer to the end devices and, as a result, reduce the latency of the resistance solution.

This research provides a novel Federated Deep Learning (FDR) framework based on Privacy Protection to take on the aforementioned problems. Here is a breakdown of the main things this study has to offer:

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

- To make fog-assisted IIoT systems highly resistant to adversarial attacks of various types, we offer a novel federated adversarial deep learning architecture. By keeping sensitive information locally on client machines and sending only metadata to the cloud server, the method may be learned and tested on fog nodes.
- We demonstrate the innovation of the suggested method by creating an image classifier in which each class is given a unique identifier that is produced independently by each fog node. Which in turn prevents privacy attacks from disclosing any sensitive data of clients by blocking the admittance of the adversary to the expected class of data.
- An adversarial federated training framework is presented, which facilitates the sharing of deep learning model meta-data with the parameter cloud server without endangering local data privacy.
- Fine-grained image recognition benchmarks in simulated fog-assisted IIoT environments are used to verify the feasibility and practicability of the proposed FDR.

This work is organised as follows: In Section II, we'll talk about the relevant research. The core of the system design argument is presented in Section III. The section on the FDR's methodology is Section IV. Experimental procedures, data, interpretations, and conclusions are discussed in Section V. Section VI draws conclusions from this study and discusses directions for future research.

II. Background and Related Work

Here, we provide a brief introduction to the research on adversarial threats in IIoT settings, privacy-preserving federated intelligence, and fog computing in IIoT settings.

A. IoT-targeted Adversarial Attacks

The results that deep learning has provided for handling IoT data are very encouraging. It still has problems when training data is scarce or unequal and can be susceptible to adversarial attacks. Adversarial instances and attacks have emerged as a powerful method for assessing deep neural networks' theoretical properties and actual efficacy in recent years. The Fast Gradient Sign Method (FGSM), which aims to insert adversarial perturbations on the road to gradients loss, is a good example of a white-box assault. [15] This method is an example of a conventional white-box attack. Another illustration of this is the Basic Iterative Method (BIM), which is described in [16] and employs the FGSM algorithm repeatedly, but with a smaller step size. In addition, the Jacobian-based Saliency Map Attack (JSMA) was described in [17] in order to identify characteristics of the input that have the greatest influence on the model's output. The adversarial samples that are produced by JSMA are determined by the calculation of forward derivatives. Carlini and Wagner [18] developed an optimization task-founded attack that they called CW with the goal of reducing the total number of perturbations while simultaneously improving the effects of the attack. In addition, DeepFool was introduced in [19] as a means of producing adversarial samples through the application of geometric principles and iterative linearization. The Simple Black-box Attack (SIMBA) [20] is a new sort of active black-box attack that is distinct from the methods described in the previous paragraph. SIMBA chooses arbitrary directions to disturb the input images rather than investigating the gradient trends in the same way that FGSM does.

Various methods of resistance or prevention have been presented in the body of scholarly work as part of an ongoing effort to develop solutions that can withstand attacks from adversaries. These methodologies could be broken down into three distinct categories [3]. First, methods that optimise the gradient computing of selected models, such as, for example, ensemble diversity [4] and Jacobian Regularization [21]. Second, methods that improve the accuracy of the gradient computation. On the other hand, the performance of this kind of technique is likely to suffer when it comes to dealing with pictures of nature. Second, there are methods that attempt to purify the inputs of the model that is being targeted by utilising extra filters or auto-encoders, such as PuVAE [3] and feature squeezing [22]. However, the extra conveniences inescapably increase the load placed on the IIoT devices that are hosting them. Third, the regularisation of targeted models may be accomplished by the utilisation of data modification strategies. An example of this would be adversarial training approaches [23], which aim to achieve an effective classification model by way of training using both adversarial and genuine data samples. One example of this would be a standard adversarial training approach. However, the existing adversarial training methodologies for specific IIoT devices only focus on a select few of the many different forms of attacks that can be made against these devices. Because of this, the trained models that were created using the aforementioned methods often could not be immediately applied by devices that were deployed in an environment that is dynamic. The work [11] attempted to overcome the constraints described above by combining federated learning [24] and adversarial training in such a way as to protect against attacks coming from a variety of sources distributed IIoT applications [11]. On the other hand, this attempt was validated using fictitious benchmark datasets rather than data taken from the real world. They also did not take into account the participants'

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

right to privacy in federated learning. This work is, to the best of the author's knowledge, one of the first countable attempts to provide privacy-preserved federated learning in order to construct a resistance framework against adversarial attacks in Fog-assisted real-world IIoT applications. The work was carried out by a team of researchers from the University of Washington and the University of Washington Bothell.

B. Federated Learning for IoT

The idea of federated learning, which is a way for training DL models on data that is irregularly dispersed across the IIoT network's various locations, has garnered a significant amount of attention from researchers as a method for training DL models [25]. The necessity of training a machine learning solution using private sensitive data that is unable to aggregated into a centralised cloud server because of the inherent privacy proprietary, and related local authority necessities is the primary impetus behind FL. This impediment prevents the data from being aggregated into a cloud server. For instance, the work [26] presented a method for the safe collaboration of data that makes use of federated learning to accomplish the goal of securing the cooperation of multiple parties in the processing of private data. the work [9] present a verifiable federated learning framework in order to protect the privacy of feed-foreward network and CNNs that have been trained to recognise handwritten images. This is accomplished by ensuring that the encrypted gradients of different participants do not become inverted. Reinforcement learning (RL) was used in [27] to recognise the heterogeneity of IoT systems depending on rating feedback. This helps the federated learning to iteratively obtain higher performance for majority of the participants. Likewise, in [28], the RL was used for collaborative inference of the DNN industrial IoT system by recasting the channel variation as a constrained Markov decision process. This was done in order to make better use of the available data. A federated learning-based protection system that makes use of security expertise was presented by the work [11] in order to withstand adversarial attacks from a variety of sources in an environment that is cloud-based and uses IIoT.

While sensitive information is being handled in IoT networks, protecting users' privacy has been highlighted as an effective strategy for preventing data from falling into the hands of unauthorised users. It is necessary for federated learning paradigms to maintain participants' privacy in order to prevent any unlawful access to sensitive data or local settings used by the participants. In this context, approaches for protecting privacy can be classified into any one of the following four categories: homomorphic encryptions (HEs), blockchain, secure multi-party computation (SMPC), and differential privacy (DP). The work [29] presented a new privacy-preserving method that protects the location information of remote distributed clients by utilising both HEs and random variation mechanism for preserving the confidentiality of the clients' location information. In order to eliminate large quantities of ciphertext processing on resource-constrained edge devices, The work [30] attempted to encrypt the model gradient rather than the local data. This was done to achieve their goal. In contrast, The work [31] presented blockchain-enabled federated learning as a means of replacing the centralised authority with a blockchain that was established with decentralised privacy standards and was designed expressly for the purpose. A similar attempt was presented in [32], in which blockchain was used to offer a decentralised incentive method to make the federated learning scheme resistant to poisoning attacks while maintaining improved authentications and participant selections. This was accomplished while keeping the participants' selections intact and maintaining improved authentications. However, despite the fact that solutions based on encryption can safeguard the privacy of local parameters and data, they frequently exhibit high computational costs, which limits their application for time-critical applications involving the Internet of Things. As a result, the researchers are moving toward using DP and SMPC methodologies in order to attain a better balance between efficacy and privacy. The goal of differential privacy is to protect the confidentiality of local information by introducing a certain amount of noise into the system in order to hide it from an adversary. For example, The work [33] integrated gaussian-based DP into the FL scheme to protect the privacy of locally trained models. They did this by employing a strategy for a random update that aimed to get rid of the attacks that were initiated by central attackers. This allowed them to defend the privacy of locally trained models. The work [34] address the privacy of the fog-based IoT network by merging the DP with a combination of blinding and HE for the purpose of protecting the model from various attacks while maintaining the privacy of the local data. This was done in order to protect the privacy of the fog-based IoT network. In addition, a recent study project that was reported in [35] had proven that merging the aforementioned categories of privacy-preservation strategies to reach a complete degree of privacy. This research was conducted in the past several years. Nevertheless, the DP-based approaches frequently suffer from severe privacy-efficiency trade-offs.In addition, The work [36] developed an SMPC technique with the purpose of designing a privacy-preserving FL system. This technique primarily makes use of a masking strategy to protect information private while it is being transmitted between participants. The work [37] presented two efficient distributed SMPC protocols with the intention of safeguarding the confidentiality of machine translation operations. The research papers that have been discussed have, without a doubt, placed an emphasis on the safety of federated learning; nonetheless, these studies see the server as

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

an enemy while ignoring adversarial attacks that originate from training participants. They also place an emphasis on the confidentiality of the data or models, despite the fact that this may be incorrect in practise. As a result of this, a Fog-Assisted FDR framework has been developed in order to make federated learning robust and resistant to adversarial attacks in an environment containing IIoT devices.

III. METHODOLOGY

Federated DL models are unable to detect adversarial samples. The primary objective of this research is to provide a novel federated learning framework with the capability of safely inferring several classes of images in edge- and fog-enabled smart cities. To solve this issue, we include malicious instances in the training dataset of the discriminatory model and tag them as attacks before retraining the model. Learning and generation are continued until the DL model reliably identifies adversarial instances. Following the convergence of the deep network, the most robust architecture against input disturbance is obtained. The noise vector is passed to the generator to engender samples with low-level and high-level components in which various preferences could be investigated to create different image components such as Gaussian process, interpolation, regression, etc. To optimize its projected final reward, the GAN selected a building in which generator G provides a likelihood of transition between states determined by Q.

$$J(\theta) = \sum_{a_j \in A} G_\theta\left(s_{i+1,j} \middle| s_i\right) \cdot Q_{D\phi}^{G_\theta}(s_i, a_j) \qquad (1)$$

The second term denotes the action-value operation attained by the discriminator's Monte Carlo search. Followingly, the proposed framework presents a knowledge distillation to empower the adversarial training of GAN under federated settings. This distillation approach can be regarded as a defensive way to improve the robustness of the deep networks for different image recognition tasks. Specifically, the first phase in our framework focuses on the training of the teacher model using a superior temperature, T, constraint to reduce the softmax possibility at the end of the federated classifier. The following formula provides a mathematical expression of the above operations:

$$p_{softmax}(z, T) = \frac{e^{z/T}}{\sum_{i=1}^{n} e^{z_{(i)}/T}} \qquad (2)$$

In the above formula, the symbol n denotes the count of labels. the symbol z denotes the output of the deep classifier such that:

$$z = \boldsymbol{W}_n \cdot \boldsymbol{a}_{n-1} + b_n \qquad (3)$$

whereas the W_n represent the matrix of learning weights, and $\boldsymbol{a}_{n-1}$ represent the activation function of the model's output. Then, the output probabilities are exploited to train the student network using small temperature constraints. The student network is trained to optimize the following objective:

$$\begin{aligned} \mathcal{L}_{student}(T) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n} \boldsymbol{y}_{ij} \cdot log\ p_{softmax}(z_{ij}, T) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n} \boldsymbol{y}_{ij} \cdot log\ \frac{e^{z_{ij}/T}}{\sum_{i=1}^{n} e^{z_{ij}/T}} \end{aligned} \qquad (4)$$

In the above formula, the symbol N represents the size of the training set, $\boldsymbol{y}_{ij}$ denote the label of the training sample, and z_{ij} denote the logit. The teacher network is trained to optimize the following objective:

$$\mathcal{L}_{teacher}(T) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n} \boldsymbol{y}_{ij} \cdot log\ \frac{e^{z_{ij}/T}}{\sum_{i=1}^{n} e^{z_{ij}/T}} \qquad (5)$$

The adversarial distillation of GAN is a method that can augment the robustness of the federated classifiers, which perform the learning using smooth targets supplied by the teacher network. This distillation mechanism comprised two primary steps. First, the learning of the teacher network. Second, the distillation of learned knowledge from a teacher to a student. The input image can be separated into high-rate components (i.e., image edges) and low-frequency components, like pixel values of color units. To capitalize on the retrieval of image representations, it is needed to estimate the cost function at both level of components. For low-frequency components, pixel loss between color units is regarded as the best way to optimize the underlying classification. Minimizing the pixel loss can be effectively

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

minimized by L_1 and L_2; however, L_2 could reach rapid convergence. On the other hand, L_1 loss can provide an ideal way to compute loss for image retrieval activities that incorporate a concurrent calculation of high-frequency components. Thus, pixel loss is calculated using L_1 as formulated below.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1] \quad (6)$$

As shown, L_1 compute the cost value for every pair of pixels from the actual image y and the engendered image $G(x,z)$. Moreover, the texture feature is a useful principle for the retrieval of high-level representations from the input image. The earliest *Pix2pix* technique suggested the formation of *patchGAN* (again) for the D to accurately criticize the image to some extent. In particular, the pGAN separates them into a set of patches, decides on the reality or incorrectness of every patch independently, and definitively computes the mean value. This computation could be considered a kind of texture loss. In our framework, we propose to improve the conditional loss of GAN, such that the excellence of the engendered image cannot be identified by the D. To this end, the cost function can be formulated as follows:

$$\mathcal{L}_{cond}(G,D) = \mathbb{E}_{x,y}[logD(x,y)] + \mathbb{E}_{x,z}\left[log\left(1 - D\big(x, G(x,z)\big)\right)\right] \quad (7)$$

Given the above two losses, the final loss function of our model can be formulated as follows:

$$G^* = \min_G \max_D \mathcal{L}_{cond}(G,D) + \lambda_1 \mathcal{L}_{L1}(G) \quad (8)$$

IV. Experiments and Analysis

A. Experimental Design

The implementation of models is performed using Pytorch library running on Python 3.7 environment is adopted during the course of the learning process. Overall software tools are installed on a Dell workstation operated with Windows 10 64-bit OS and are armed with Intel (R) Xeon (R) CPU E5-2670 0@ 2.60GHz, a GPU made by NVIDIA Tesla server, and a memory that is 128 GB.

B. Evaluation measures

The evaluation metrics are calculated using the following equations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \times 100 \quad (11)$$

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (12)$$

C. Dataset Description

Two well-known fine-grained image classification datasets are employed to evaluate the performance of the proposed FDR. Food-101[2]: the data comprise 101 categories of food, each category comprises 750 training images as well as 250 testing images resulting in a total of 101k images. all the images are rescaled to have unified dimensions of 64×64. CIFAR-100 [3]: the data consist of 60,000 RGB images with the size of 32×32, which belongs to 100 distinct classes.

The Shuffle-Net [4] is employed as a deep learning model for image recognition owing to its lightweight nature. A batch normalization layer is employed after the fixed layer for stability purposes [5]. Since the data is large enough and properly distributed among classes, the training process did not involve any kind of data augmentation. The test set is used as held-out data to assess the generalizability of the model. A grid search algorithm is employed to find the optimal hyperparameters for the proposed FDR framework based on stratified 5-fold cross-validation and according to validation accuracy, then, the corresponding results are shown in Table I.

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

Table I: The grid search hyperparameters for the FDR *framework.*

Hyperparameters	Search Interval	Values
Optimizer	['SGD', 'RMSprop', 'Adagrad', 'Adam', 'Nadam']	Adam
Learning rate	[0.0005, 0.001, 0.005, 0.01, 0.05, 0. 1]	0.005
Batch-size	[32, 64, 96, 128, 192, 264]	128
Weight initialization	['uniform', 'lecun_uniform', 'normal', 'zero', 'glorot_normal', 'glorot_uniform', 'he_normal', 'he_uniform']	'glorot_ uniform' [6]
Training epochs	[40, 50, 60, 70, 80, 90, 100]	60

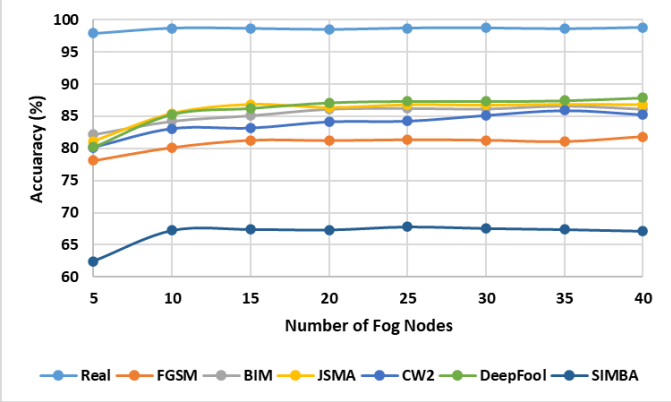

Figure. 1: Performance of the FDR with respect to various number of nodes (Food-101 dataset).

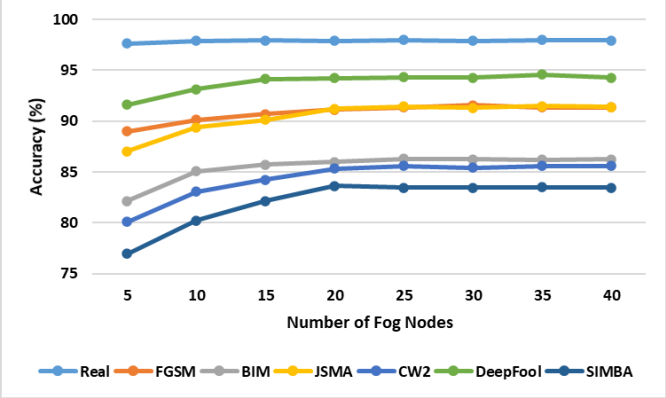

Figure 2: Performance of the FDR with respect to various number of nodes (CIFAR-100 ***dataset).***

Additional hyperparameters include the number of communication rounds and identifier length, which is set to be 20 and 512, respectively. In all experiments, the beforementioned is set by default and only the changed hyperparameters are discussed in its relevant section.

D. Results

Table II: Classification accuracy for different resistance techniques against various Adversarial Attacks on Food-101 Dataset

		Resistance Approaches								
		None	**Adv (FGSM)**	**Adv (BIM)**	**Adv (JSMA)**	**Adv (CW2)**	**Adv (DeepFool)**	**Adv (All)**	**FDA3**	FDR
Attacks	Real	97.79	97.72	95.53	96.21	96.64	96.72	97.71	97.81	98.7
	FGSM	30.16	70.28	67.2	63.57	62.76	64.57	75.61	77.25	80.14
	BIM	28.75	72.84	71.53	71.64	69.82	70.08	80.13	79.25	84.25
	JSMA	2.16	58.37	57.25	71.11	72.16	68.57	82.53	80.78	85.46
	CW2	0.2	44.13	46.71	70.25	72.13	46.13	81.13	78.24	83.09
	DeepFool	1.46	62.45	61.36	60.97	61.25	63.41	83.74	82.94	85.31
	SIMBA	23.41	**62.31**	62.02	59.17	60.45	61.34	65.43	64.67	67.25

Table III: Classification accuracy for different resistance techniques against various Adversarial Attacks on CIFAR-100 Dataset

		Resistance Approaches								
		None	**Adv (FGSM)**	**Adv (BIM)**	**Adv (JSMA)**	**Adv (CW2)**	**Adv (DeepFool)**	**Adv (All)**	**FDA3**	FDR
Attacks	Real	98.13	97.72	97.25	96.99	97.03	96.72	98.01	97.25	97.87
	FGSM	36.27	73.45	67.2	66.87	70.13	69.87	80.03	88.07	90.12
	BIM	35.71	74.25	71.53	76.23	67.33	76.34	82.45	82.67	85.09
	JSMA	11.14	66.15	57.25	74.98	74.31	73.15	83.64	86.47	89.37
	CW2	3.4	52.74	46.71	80.03	75.13	55.67	81.13	78.24	83.09
	DeepFool	7.6	70.02	61.36	68.78	80.07	69.89	86.23	90.13	93.14
	SIMBA	19.18	**70.94**	62.02	61.74	71.28	73.48	73.14	75.67	80.25

In the comparative experiments, ten fog nodes participate to the federated training with single cloud server as a parameter or aggregation server. Each participant is assumed that to have 100 real samples for adversarial training.

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

Six popular categories of adversarial attacks are considered in our comparative experiments including FGSM, BIM, JSMA, CW, DeepFool, and SIMBA. Whereas each category was employed to attack three out of ten participants. On The Way To accommodate the adversarial training, each participant generated 100 adversarial samples for the 100 real samples utilizing the designated attack method, correspondingly. It's important to note that all of the adversarial examples here are produced using transference attacks, which assume that the initial model could be acquired but the intermediate retrained models were inaccessible to malicious actors. Like [17], the hyperparameter was assigned a value of 0.5 to promote both regular and adversarial costs contribute evenly to the final loss. To use adversarial attacks in federated training, since there are 100 couples of real and adversarial samples on each participant, the epoch size is set to be 60 (based on convergence), wherever each epoch is employed to retrain all the accumulated couples at each participant.

To allow the comparing the performance of the proposed FDR framework with conventional adversarial resistance approaches. the model at each participant is updated by retraining it locally with 100 couples of samples using different categories of attacks independently. The term None represents the original model without any retraining process. The term $Adv(A)$, where $A \in \{$ FGSM, BIM, JSMA, CW2, DeepFool, SIMBA $\}$ to imply the classification model of each participant retrained locally based on the 100 adversarial samples initiated with attacks of category A. In addition to conventional approaches, the proposed FDR framework is compared with the recent proposed federated defense. Furthermore, the comparison also considers the case where six attack categories are applied on each participant node. This means that every node has a total of 500 adversarial samples for local retraining, this case is termed as $Adv(All)$. In Table II and Table III, the classification accuracy and F1-measure are reported for each competent approach against different categories of adversarial attacks on the test set of the Food-101 dataset. It is worth noting that all participating fog nodes exhibit similar performance against different attacks. Thus, the tabulated results are randomly collected from an arbitrarily chosen participant among all the ten participating nodes. According to the tabulated results, it could be noted that the proposed FDR achieves the best recognition performance among other competing approaches except for the real scenario. For the real test set (i.e., no adversarial), the none approach marginally surpasses the proposed FDR by 0.37% and 0.41% on accuracy and F1-measure, respectively. This can be justified by the fact that the proposed FDR contains some adversarial samples through the retraining process. On other hand, the proposed FDR overcomes the competent approaches on other attack scenarios with great accuracy improvements (FGSM: 2.89, BIM: 5, JSMA:4.68, CW: 4.85, DeepFool: 2.37, SIMBA: 2.58) as well as F1-measure improvements (FGSM: 2.95, BIM: 5.06, JSMA: 4.74, CW: 4.91, DeepFool: 2.43, SIMBA: 2.64). More importantly, the proposed FDR is the competent federated defense method. This further explains the ability of the proposed FDR in preventing participants from accessing the data of each other's maintaining the privacy of their data and keeping resistance against different kinds of adversarial attacks.

The previous experiments only explore the IIoT federated application that trained ten fog participants. Nevertheless, real-world IIoT applications often include dozes of or thousands of nodes. Thus, to validate whether the proposed may be used for a broad range of IIoT applications, stability experiments are performed to investigate the scalability of the

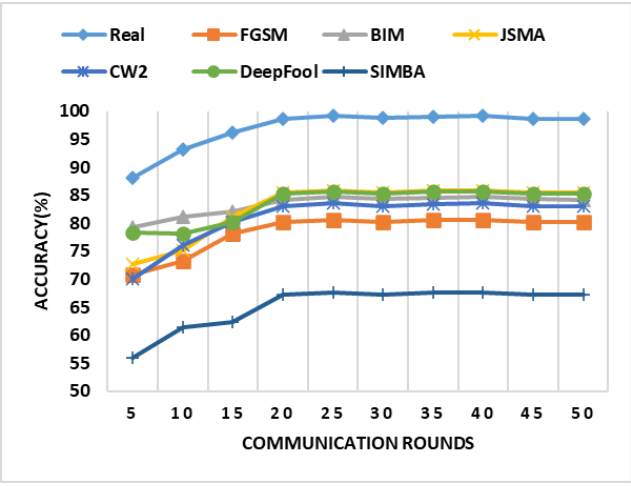

Figure 3: Ablation analysis for of number of communication rounds on Food-101dataset.

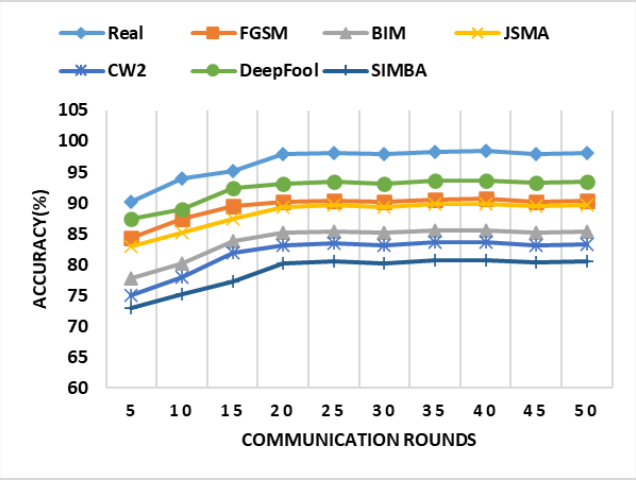

Figure 4: Analyzing the impact of number of communication rounds on classification performance on CIFAR-100 dataset

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

proposed FDR. Fig. 1 and Fig. 2 show the classification accuracy along with the increase of the number of fog nodes over the CIFAR-100 and Food-101 dataset respectively. In this experiment, the proposed FDR is experimented using a different number of fog nodes from 5 to 40 fog nodes. As employed in the previous experiments, six categories of adversarial attacks namely FGSM, BIM, JSMA, CW, Deep Fool, and SIMBA for sample generation. In these experiments, three participants out of ten are assumed to be under attack. The stability experiments consider seven categories of test instances, where real represent the case where the test set has no attack samples. Typically, it could be seen that image recognition accuracy is top in case of real scenario. Additionally, we discover that the classification marginally improves and begins to stabilise until the number of fog nodes reaches around 15 or 20 participants in federated training. The remaining six hostile test sets all show the same pattern. For the FGSM test set, increasing the sample size from 15 to 40 improves accuracy from 81.26% to 81.83%. it worth noting that in the federated training does not consider SIMBA attack, thus, the classification performance of the SIMBA test set is the smallest. Nevertheless, the accuracy improvement could be observed by increasing the number of fog nodes to 20 in the case of the CIFAR-100 dataset.

In order to validate the selection of the number of communication rounds and understand the behavior of the proposed framework through federated training rounds, and additional experiments are performed by evaluating the performance of the proposed framework under different communication rounds. The corresponding results for Food-101 and CIFAR-100 datasets are presented in Fig. 3 and Fig. 4, respectively. surprisingly, the proposed start after 15 or 20 rounds for all attack classes on both datasets. This further explains the smooth convergence ability of the proposed framework and thereby indicates a small communication overhead during training.

V. CONCLUSIONS

This study presents a Privacy Protected Federated Resistance framework called FDR, which seek to enable the deep learning model to resist against broad range adversarial attacks in Fog-assisted IoT applications. In particular, DL-based image recognition was deployed on the fog nodes, where arbitrary class identifiers are generated to denote the hosted class labels. This identifier-based technique enables reliable and efficient learning on the fog nodes by using high dimensional identifiers, where class weights are safeguarded from any operational adversary that might try to initiate an adversarial attack. The proposed FDR framework promotes realizing the best recognition performance, rapid convergence, while preserving the data privacy. The effectiveness and proficiency of the proposed method are validated through experimental and theoretical evaluations.

REFERENCES

[1] M. B. Sariyildiz, R. G. Cinbis, and E. Ayday, "Key protected classification for collaborative learning," *Pattern Recognit.*, vol. 104, 2020, doi: 10.1016/j.patcog.2020.107327.

[2] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 - Mining discriminative components with random forests," 2014, doi: 10.1007/978-3-319-10599-4_29.

[3] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images.(2009)," *Cs.Toronto.Edu*, pp. 1–58, 2009.

[4] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet V2: Practical guidelines for efficient cnn architecture design," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11218 LNCS. pp. 122–138, 2018, doi: 10.1007/978-3-030-01264-9_8.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, vol. 1, pp. 448–456.

[6] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Journal of Machine Learning Research*, 2010, vol. 9, pp. 249–256.

[7]. X. Zhang, Y. Zhou, S. Pei, J. Zhuge and J. Chen, "Adversarial Examples Detection for XSS Attacks Based on Generative Adversarial Networks," in *IEEE Access*, vol. 8, pp. 10989-10996, 2020, doi: 10.1109/ACCESS.2020.2965184.

[8]. K. Madono, M. Tanaka, M. Onishi and T. Ogawa, "SIA-GAN: Scrambling Inversion Attack Using Generative Adversarial Network," in *IEEE Access*, vol. 9, pp. 129385-129393, 2021, doi: 10.1109/ACCESS.2021.3112684.

[9]. D. Wang, L. Dong, R. Wang, D. Yan and J. Wang, "Targeted Speech Adversarial Example Generation With

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

Generative Adversarial Network," in *IEEE Access*, vol. 8, pp. 124503-124513, 2020, doi: 10.1109/ACCESS.2020.3006130.
[10]. V. R. Kebande, S. Alawadi, F. M. Awaysheh and J. A. Persson, "Active Machine Learning Adversarial Attack Detection in the User Feedback Process," in *IEEE Access*, vol. 9, pp. 36908-36923, 2021, doi: 10.1109/ACCESS.2021.3063002.
[11]. X. Hu, D. Cheng, J. Chen, X. Jin and B. Wu, "Multiontology Construction and Application of Threat Model Based on Adversarial Attack and Defense Under ISO/IEC 27032," in *IEEE Access*, vol. 10, pp. 117955-117972, 2022, doi: 10.1109/ACCESS.2022.3220637.
[12]. A. Kuppa and N. -A. Le-Khac, "Adversarial XAI Methods in Cybersecurity," in *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4924-4938, 2021, doi: 10.1109/TIFS.2021.3117075.
[13]. Y. -Y. Chen, C. -T. Chen, C. -Y. Sang, Y. -C. Yang and S. -H. Huang, "Adversarial Attacks Against Reinforcement Learning-Based Portfolio Management Strategy," in *IEEE Access*, vol. 9, pp. 50667-50685, 2021, doi: 10.1109/ACCESS.2021.3068768.
[14]. I. Aliyu, S. Van Engelenburg, M. B. Mu'Azu, J. Kim and C. G. Lim, "Statistical Detection of Adversarial Examples in Blockchain-Based Federated Forest In-Vehicle Network Intrusion Detection Systems," in *IEEE Access*, vol. 10, pp. 109366-109384, 2022, doi: 10.1109/ACCESS.2022.3212412.
[15]. I. Alsmadi *et al*., "Adversarial Machine Learning in Text Processing: A Literature Survey," in *IEEE Access*, vol. 10, pp. 17043-17077, 2022, doi: 10.1109/ACCESS.2022.3146405.
[16]. Y. Zheng, Y. Lu and S. Velipasalar, "An Effective Adversarial Attack on Person Re-Identification in Video Surveillance via Dispersion Reduction," in *IEEE Access*, vol. 8, pp. 183891-183902, 2020, doi: 10.1109/ACCESS.2020.3024149.
[17]. W. Zhang, "Generating Adversarial Examples in One Shot With Image-to-Image Translation GAN," in *IEEE Access*, vol. 7, pp. 151103-151119, 2019, doi: 10.1109/ACCESS.2019.2946461.
[18]. C. Park, Y. Kim, J. -G. Park, D. Hong and C. Seo, "Evaluating Differentially Private Generative Adversarial Networks Over Membership Inference Attack," in *IEEE Access*, vol. 9, pp. 167412-167425, 2021, doi: 10.1109/ACCESS.2021.3137278.
[19]. X. Zhang, J. Wang and S. Zhu, "Dual Generative Adversarial Networks Based Unknown Encryption Ransomware Attack Detection," in *IEEE Access*, vol. 10, pp. 900-913, 2022, doi: 10.1109/ACCESS.2021.3128024.
[20]. F. Nikfam, A. Marchisio, M. Martina and M. Shafique, "AccelAT: A Framework for Accelerating the Adversarial Training of Deep Neural Networks Through Accuracy Gradient," in *IEEE Access*, vol. 10, pp. 108997-109007, 2022, doi: 10.1109/ACCESS.2022.3213734.
[21]. Y. Sun and L. Fu, "A New Threat for Pseudorange-Based RAIM: Adversarial Attacks on GNSS Positioning," in *IEEE Access*, vol. 7, pp. 126051-126058, 2019, doi: 10.1109/ACCESS.2019.2939141.
[22]. T. -T. -H. Le, H. Kang and H. Kim, "Robust Adversarial Attack Against Explainable Deep Classification Models Based on Adversarial Images With Different Patch Sizes and Perturbation Ratios," in *IEEE Access*, vol. 9, pp. 133049-133061, 2021, doi: 10.1109/ACCESS.2021.3115764
[23]. X. Kang, B. Song, X. Du and M. Guizani, "Adversarial Attacks for Image Segmentation on Multiple Lightweight Models," in *IEEE Access*, vol. 8, pp. 31359-31370, 2020, doi: 10.1109/ACCESS.2020.2973069.
[24]. X. Zhang, Y. Zhou, S. Pei, J. Zhuge and J. Chen, "Adversarial Examples Detection for XSS Attacks Based on Generative Adversarial Networks," in *IEEE Access*, vol. 8, pp. 10989-10996, 2020, doi: 10.1109/ACCESS.2020.2965184.
[25]. R. Wang, Z. Chen, H. Dong and Q. Xuan, "You Can't Fool All the Models: Detect Adversarial Samples via Pruning Models," in *IEEE Access*, vol. 9, pp. 163780-163790, 2021, doi: 10.1109/ACCESS.2021.3133334.
[26]. K. Yamanaka, R. Matsumoto, K. Takahashi and T. Fujii, "Adversarial Patch Attacks on Monocular Depth Estimation Networks," in *IEEE Access*, vol. 8, pp. 179094-179104, 2020, doi: 10.1109/ACCESS.2020.3027372.
[27]. Z. Li, C. Feng, J. Zheng, M. Wu and H. Yu, "Towards Adversarial Robustness via Feature Matching," in *IEEE Access*, vol. 8, pp. 88594-88603, 2020, doi: 10.1109/ACCESS.2020.2993304.
[28]. Á. L. Perales Gómez, L. F. Maimó, F. J. G. Clemente, J. A. M. Morales, A. H. Celdrán and G. Bovet, "A Methodology for Evaluating the Robustness of Anomaly Detectors to Adversarial Attacks in Industrial Scenarios," in *IEEE Access*, vol. 10, pp. 124582-124594, 2022, doi: 10.1109/ACCESS.2022.3224930.
[29]. Y. Bakhti, S. A. Fezza, W. Hamidouche and O. Déforges, "DDSA: A Defense Against Adversarial Attacks Using Deep Denoising Sparse Autoencoder," in *IEEE Access*, vol. 7, pp. 160397-160407, 2019, doi: 10.1109/ACCESS.2019.2951526.
[30]. F. O. Catak, M. Kuzlu, E. Catak, U. Cali and O. Guler, "Defensive Distillation-Based Adversarial Attack

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022

Mitigation Method for Channel Estimation Using Deep Learning Models in Next-Generation Wireless Networks," in *IEEE Access*, vol. 10, pp. 98191-98203, 2022, doi: 10.1109/ACCESS.2022.3206385.

[31]. R. H. Randhawa, N. Aslam, M. Alauthman, H. Rafiq and F. Comeau, "Security Hardening of Botnet Detectors Using Generative Adversarial Networks," in *IEEE Access*, vol. 9, pp. 78276-78292, 2021, doi: 10.1109/ACCESS.2021.3083421.

[32]. Z. Liu and X. Yin, "LSTM-CGAN: Towards Generating Low-Rate DDoS Adversarial Samples for Blockchain-Based Wireless Network Detection Models," in *IEEE Access*, vol. 9, pp. 22616-22625, 2021, doi: 10.1109/ACCESS.2021.3056482.

[33]. X. Kuang, H. Liu, Y. Wang, Q. Zhang, Q. Zhang and J. Zheng, "A CMA-ES-Based Adversarial Attack on Black-Box Deep Neural Networks," in *IEEE Access*, vol. 7, pp. 172938-172947, 2019, doi: 10.1109/ACCESS.2019.2956553.

DOI: https://doi.org/10.54216/JISIoT.060205
Received: March 18, 2022 Accepted: July 28, 2022