CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Conference Paper

Poisoning Federated Learning with Graph Neural Networks in Internet of Drones

Kai Li*
Alam Noor*
Wei Ni
Eduardo Tovar*
Xiaoming Fu
Ozgur B. Akan

*CISTER Research Centre

CISTER-TR-240501

2024/07/29

Poisoning Federated Learning with Graph Neural Networks in Internet of Drones

Kai Li*, Alam Noor*, Wei Ni, Eduardo Tovar*, Xiaoming Fu, Ozgur B. Akan

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: kai@isep.ipp.pt, alamn@isep.ipp.pt, Wei.Ni@data61.csiro.au, emt@isep.ipp.pt, fu@cs.uni-goettingen.de, oba21@cam.ac.uk

https://www.cister-labs.pt

Abstract

Internet of Drones (IoD) is an innovative technology that integrates mobile computing capabilities with drones, enabling them to process data at or near the location where it is collected. Federated learning can significantly enhance the efficiency and effectiveness of data processing and decision-making in IoD. Since federated learning relies on aggregating updates from multiple drones, a malicious drone can generate poisoning local model updates that involves erroneous information, leading to incorrect decisions or even dangerous situations. In this paper, a new data-independent model poisoning attack is developed to manipulate federated learning accuracy, which does not rely on training data at drones. The proposed attack leverages an adversarial graph neural network (A-GNN) to generate poisoning local model updates based on the benign local models overheard. Particularly, the A-GNN discerns the graph structural correlations between the benign local models and the features of the training data that underpin these models. The graph structural correlations are reconstructively manipulated at the malicious drone to crafts poisoning local model updates, where the training loss of the federated learning is maximized.

Poisoning Federated Learning with Graph Neural Networks in Internet of Drones

Kai Li*§, Alam Noor*, Wei Ni‖, Eduardo Tovar*, Xiaoming Fu†‡, Ozgur Akan§¶
*CISTER Research Centre, Portugal
Email: {kai, alamn, emt}@isep.ipp.pt
†Center for Internet & Society, Fudan University, China
‡Institute of Computer Science, University of Göttingen, Germany
Email: fu@ieee.org
§Internet of Everything Group, University of Cambridge, UK
¶Center for NeXt-Generation Communications (CXC), Koç University, Turkey
Email: {kl596, oba21}@cam.ac.uk
‖CSIRO, Australia
Email: wei.ni@data61.csiro.au

***Abstract*—Internet of Drones (IoD) is an innovative technology that integrates mobile computing capabilities with drones, enabling them to process data at or near the location where it is collected. Federated learning can significantly enhance the efficiency and effectiveness of data processing and decision-making in IoD. Since federated learning relies on aggregating updates from multiple drones, a malicious drone can generate poisoning local model updates that involves erroneous information, leading to incorrect decisions or even dangerous situations. In this paper, a new data-independent model poisoning attack is developed to manipulate federated learning accuracy, which does not rely on training data at drones. The proposed attack leverages an adversarial graph neural network (A-GNN) to generate poisoning local model updates based on the benign local models overheard. Particularly, the A-GNN discerns the graph structural correlations between the benign local models and the features of the training data that underpin these models. The graph structural correlations are reconstructively manipulated at the malicious drone to crafts poisoning local model updates, where the training loss of the federated learning is maximized.**

***Index Terms*—Internet of Drones (IoD), Federated learning, Adversarial Graph Neural Networks (A-GNN), Poisoning attack, Mobile computing**

I. Introduction

Internet of Drones (IoD) is an innovative technology that integrates mobile computing capabilities with drones, enabling them to process data at or near the location where it is collected [1]. By combining the mobility and flexibility of drones with the power of mobile computing, this approach allows for real-time data analysis and decision-making in the air, significantly reducing the latency and bandwidth issues associated with transmitting data to distant cloud servers [2]. This is particularly beneficial for applications requiring immediate response and action, such as disaster response, environmental monitoring, and urban planning [3]. IoD transforms how we collect, process, and leverage data from the skies, opening new possibilities for more efficient, responsive, and intelligent systems [4].

Federated learning can significantly enhance the efficiency and effectiveness of data processing and decision-making [5]. Fig. 1 depicts federated learning in an IoD scenario, where each drone can be equipped with sensors and computing units, collecting data from its environment. Instead of sending all this raw data back to a central server, which can be bandwidth-intensive and potentially compromise privacy, each drone processes the data locally to update a shared machine-learning model. These local model updates are the only information exchanged between the drones and a central server or amongst themselves. Since federated learning relies on aggregating updates from multiple drones, a malicious drone's poisoning local model updates can introduce erroneous information, leading to incorrect decisions or even dangerous situations [6]. This vulnerability is particularly critical in scenarios like disaster response, environmental monitoring, or urban surveillance, where the accuracy and reliability of data are paramount.

In this paper, we develop an innovative approach for model poisoning attacks to manipulate federated learning accuracy in IoD. Specifically, the malicious drone leverages an adversarial graph neural network (A-GNN), which is capable of creating poisoning local model updates by analyzing and utilizing the characteristics of benign local and global models. A malicious drone discreetly eavesdrops the local models shared by benign drones and the global model disseminated by the server. The A-GNN excels in identifying and interpreting complex patterns and structures found in data represented as graphs. Its proficiency lies in compressing graph data into a compact, lower-dimensional space, while maintaining the graph's critical topological attributes. The malicious drone then reconstructively modifies the graph's structure, aiming to preserve the local models' structural traits and to increase the training loss in federated learning. Subsequently, the malicious drone crafts detrimental local models based on this altered graph structure, aligning them with the data characteristics of

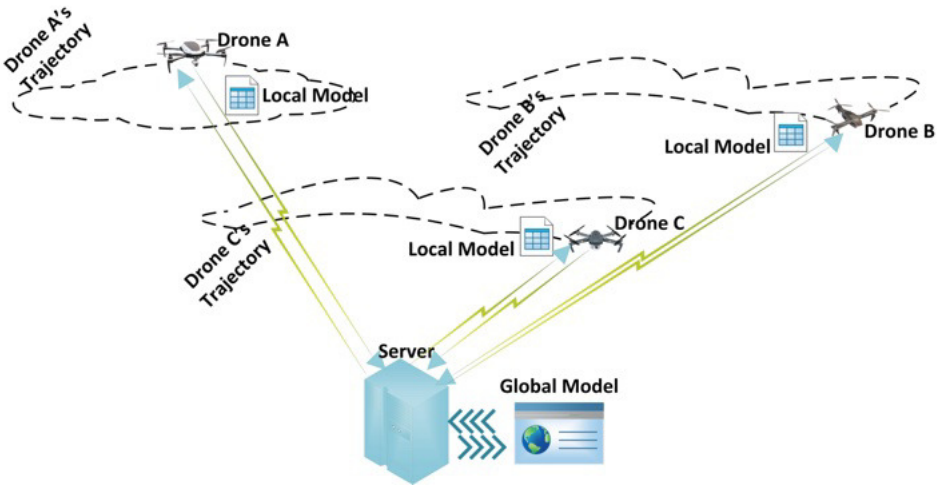


Fig. 1: Federated learning-enabled IoD, where the drone can be equipped with sensors and computing units, collecting data from its environment. A machine learning model is trained at the drone to produce a local model update. The server aggregates the local model updates from the drones to train a global model.

the benign local models. Consequently, the poisoning local models can significantly undermine the integrity of the global model, while maintaining compatibility of the poisoning local model with the benign ones, which makes the proposed A-GNN attack difficult to detect.

This paper presents several key contributions, including:

- The development of a new cyberattack for creating data-independent, poisoning local models with the malicious drone to minimize federated learning accuracy in IoD. This design alters the correlations found in benign local models while preserving their authentic data characteristics.
- The investigation of a new A-GNN attack. This framework is trained in conjunction with sub-gradient descent techniques to deceptively regenerate the correlations within the local models. This is achieved while ensuring that the poisoning local models remain undetected.
- The proposed A-GNN attack was implemented on a Support Vector Machine (SVM) model, utilizing the PyTorch framework version 1.12.1 and Python version 3.9.12. Based on MNIST datasets, performance shows that the A-GNN attack markedly undermines federated learning efficiency in IoD. This is evidenced by a notable decrease in training accuracy, which dropped below 50% on benign drones.

This paper is structured in the following manner: Section II provides an overview of adversarial attacks and defense models in IoD and federated learning. In Section III, we explore system model of federated learning-enabled IoD in the context of flight trajectories, communication channel, along with a model of eavesdropping. The design of our A-GNN attack is detailed in Section IV. We present our performance evaluation in Section V. The paper is concluded in Section VI.

II. Related Work

In this section, we present the related work in terms of the adversarial attacks and defense models in IoD and federated learning.

A semidefinite relaxation framework aimed at identifying and countering spoofing attacks on drones is studied in [7]. The detection of malicious drones is formulated as a problem of localization feasibility, utilizing both reported positions and distance measurements. The semidefinite relaxation framework transforms the inherently non-convex problem of localization into a manageable semidefinite program, which capitalizes on the spatial proximity of neighboring drones to pinpoint and isolate the malicious drones. Since offloaded information from drones to a server could be compromised by an eavesdropper, an IoD with energy harvesting is designed in [8]. A secure and energy-efficient computational offloading model is developed for improving the confidentiality of the computational offloading from drones, which generates noise signals to hinder eavesdropping by malicious drones. The authors in [9] introduces an IoD where a legitimate drone is deployed to monitor the flight of malicious drones, aiming to mitigate safety and security risks. To obtain flight data from the malicious drones, the legitimate drone employs a tactical eavesdropping and jamming approach which intentionally disrupts the malicious drones' communication, compelling it to lower its data transmission rate, in turn, increasing the likelihood of successful eavesdropping. A tracking algorithm is also developed for the legitimate drone to utilize the data from the eavesdropped packets, along with the angle-of-arrival and the received signal strength indicators from the malicious drone.

The authors in [10] focus on categorization of federated learning threats. Based on the specific goals, poisoning at-

tacks can be divided into two types: untargeted and targeted poisoning attacks. Untargeted attacks aim to degrade the overall performance of the system, while targeted attacks focus on manipulating the system to produce specific erroneous outcomes. A overview of poisoning attacks and corresponding defense strategies in federated learning is provided in [11], where existing poisoning attacks are categorized based on the implementation methods and objectives. The defense strategies for federated learning are classified into three categories: model analysis, which involves scrutinizing the models for signs of tampering; Byzantine robust aggregation, focusing on resilient aggregation methods to counteract the influence of malicious models; and verification-based strategies, where the emphasis is on verifying the integrity and authenticity of the models before they are aggregated. A collusive model poisoning attack on federated learning is presented in [12], which allows malicious participants to create untargeted poisoning local models that adhere to specific distance constraints. The collusion-based attack generates the malicious local models aiming to reduce the convergence and accuracy of the global model. A model poisoning attack is designed in [13], which injects adversarial neurons into the redundant spaces of a neural network to generate the malicious local model. The redundant neurons play a crucial role in facilitating the poisoning attack, yet they exhibit minimal correlation with the primary task of the federated learning. As a result, the model poisoning attack is designed in such a way that it does not compromise the performance of the main task on the shared global model. A defense framework is introduced to guard against poisoning attacks in federated learning systems [14]. The defense framework features a proof generation method that enables participants to produce verifiable proofs, determining whether their contributions are malicious. An aggregation rule is designed to maintain the training accuracy in the global model.

The poisoning attacks in the literature targeting federated learning systems fall short in accounting for the subtle interconnections among various local model updates. This oversight can be identified by poisoning defense frameworks that measure Euclidean distances among the local model updates. Unlike existing approaches, the proposed A-GNN attack in IoD introduced in this paper represents a novel technique for model poisoning that operates independently of the data itself. This method specifically targets the manipulation of correlations between various data features in carefully chosen benign local models. At the same time, the A-GNN attack maintains the authenticity of the data features underlying these models. This strategy ensures that the malicious drone's poisoning local models remain undetectable.

III. System Model of Federated Learning-enabled IoD

In this section, we study a federated learning training process in IoD. A malicious drone generates and uploads poisoning local model updates with the aim of gradually poisoning the global model.

A. Local model training at the drone

There are I benign drones along with one authorized yet malicious drone. For each benign drone i ($\in [1, I]$), it possesses a dataset of size D_i. The variables x_j^i and y_j^i represent, respectively, the input from captured data and the output from the federated learning model on drone i, with j ranging from 1 to D_i [15]. The training loss function for drone i is represented as $f_j(w_i; x_j^i, y_j^i)$, which quantifies the approximation errors based on the input x_j^i and output y_j^i. The term w_i indicates the weight parameter of the loss function in the federated learning model under training. For example, this function could adopt a linear regression model, such as $f_j(w_i; x_j^i, y_j^i) = \frac{1}{2}(w_i^T x_j^i - y_j^i)^2$, or a logistic regression model, like $f_j(w_i; x_j^i, y_j^i) = y_j^i \log\Big(1 + \exp\big(-w_i^T x_j^i\big)\Big) - (1 - y_j^i)\log\Big(1 - \frac{1}{1+\exp\left(-w_i^T x_j^i\right)}\Big)$. The notation $(\cdot)^T$ is used to denote the transpose [16].

For drone i in each communication round of federated learning, the local loss function, given the dataset size D_i, is expressed as:

$$F_i(w_i) = \frac{1}{D_i}\sum_{j=1}^{D_i} f_j(w_i; x_j^i, y_j^i) + \alpha \mathcal{N}(w_i), \tag{1}$$

where $\mathcal{N}(\cdot)$ is a regularizer function that represents the effect of the local training noise, and $\alpha \in [0, 1]$ is a coefficient [17].

B. Flight and channel model

The position of a drone can be represented by the coordinates (U_x^i, U_y^i, U_z^i). The drone operates in an altitude hold mode, namely, the altitude U_z^i remains constant. The drone's patrol velocity is given by v_i, which is bound by the minimum and maximum permissible velocities, $V_{\min}$ and $V_{\max}$, respectively, such that $V_{\min} \le v_i \le V_{\max}$. Δt_i is the time when the drone moves from (U_x^i, U_y^i, U_z^i) to the next position. The acceleration at (U_x^i, U_y^i, U_z^i) can be calculated as follows:

$$\Delta v_i / \Delta t_i = (v_i' - v_i)/\Delta t_i, \tag{2}$$

where v_i' is the velocity at the next position and the acceleration is constrained by the equation [18]:

$$0 \le \Delta v_i / \Delta t_i \le V_{max}/\Delta t_i. \tag{3}$$

When communicating with the server, the drone's line-of-sight (LoS) probability with the ground server is defined as

$$\mathrm{Pr}_{\mathrm{LoS}} = \frac{1}{1 + a\exp(-b[\gamma_i - a])} \tag{4}$$

where a and b are parameters of the Sigmoid function [19]. γ_i represents the elevation angle between drone i and the

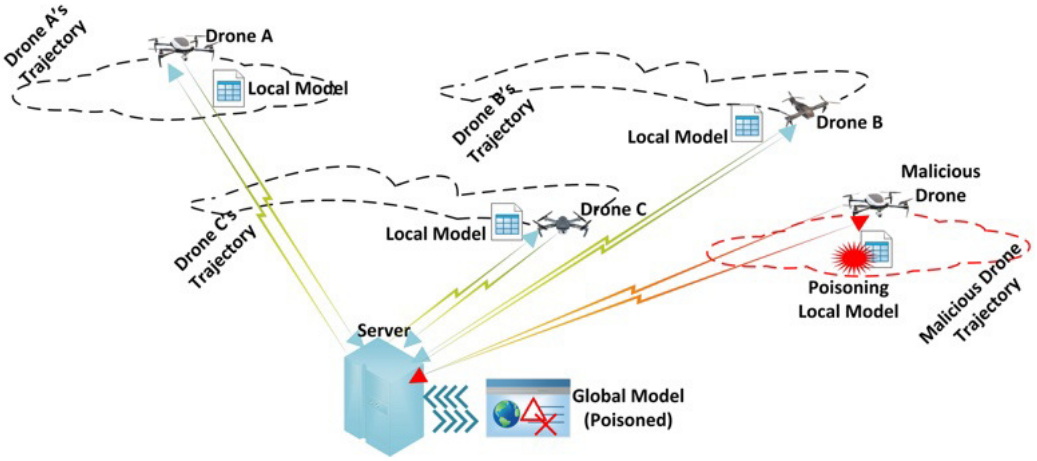


Fig. 2: Based on the benign local model updates overheard, the malicious drone conducts the A-GNN attack.

server. Additionally, the path loss between drone i and the server can be determined by

$$P_i = \mathrm{Pr}_{\mathrm{LoS}}(\gamma_i)(\phi_{\mathrm{LoS}} - \phi_{\mathrm{NoS}}) + 20\log(\mathcal{A}\gamma_i) + 20\log(\mathrm{freq}_0) + 20\log(4\pi/v_c) + \phi_{\mathrm{NoS}} \tag{5}$$

where $\mathcal{A}$ is the radius of the drone's radio coverage. freq_0 is the carrier frequency, and v_c is the speed of light. ϕ_{LoS} and ϕ_{NoS} stand for the excessive path loss of LoS and non-LoS, respectively. The terms (ϕ_{LoS}, ϕ_{NoS}) represent the excessive path loss for LoS and non-LoS conditions, respectively. The pair of values for (ϕ_{LoS}, ϕ_{NoS}) may vary as (0.1, 21), (1.0, 20), (1.6, 23), or (2.3, 34) depending on the environmental context–suburban, urban, dense urban, or highrise urban scenarios [20].

IV. The Proposed Adversarial Graph Neural Network Attack

In this section, we study the threat model where the malicious drone develops and dispatches poisoning local model updates, with the strategic aim of gradually degrading the integrity of the global model in the federated learning.

A. Threat model

Fig. 2 presents the threat model, where the malicious drone conducts the A-GNN attack based on the benign local model updates overheard. This involves crafting a poisoning local model update that is transmitted to the server. This poisoning local model is designed to manipulate the federated learning process in a contrary direction (i.e., minimizing the accuracy), resulting in the corruption of the local model updates from the benign drones. In federated learning-enabled IoD operating within wireless environments, this type of attack could be especially critical because of the inherent broadcast characteristics of radio communication [21].

The attacker could be a compromised legitimate drone or a malicious drone, with the objective of maximizing the training loss of federated learning. The malicious drone methodically generates and uploads poisoning local models, thereby incrementally poisoning the global model, denoted as w_g. This, in turn, adversely affects the local models of benign drones, i.e., w_i. In particular, w_a is used to denote the local model update of the malicious drone [22].

Without knowing the adversarial intentions of the malicious drone, the server proceeds to aggregate the local model updates from the drones. This mix includes both the benign and the poisoning local model updates, inadvertently leading to the formation of a manipulated global model, which is denoted as w_g^a. The data size is calculated by $D = \sum_{i=1}^{I} D_i + D_a$, where D_a represents the reported data size from the malicious drone. Consequently, the formulation of the manipulated global model can be given as

$$w_g^a = \sum_{i=1}^{I} \frac{D_i}{D} w_i + \frac{D_a}{D} w_a, \tag{6}$$

B. A-GNN for poisoning federated learning accuracy

According to the loss function in (1) and the manipulated global model w_g^a in (6), the poisoning local model update is generated to achieve $\max\{F(w_g^a)\}$ at the server, while the Euclidean distance between w_a and w_i ($\forall i \in [1, I]$) is below a predetermined threshold E_T. Particular, E_T is set to guarantee that the generated poisoning local model update closely resembles the benign model in Euclidean space. Reducing the Euclidean distance is crucial since the poisoning model needs to evade detection by the server's defense mechanism.

The A-GNN attack solves $w_a^\star = \arg\max\{F(w_g^a)\}$ by maintaining and strategically altering the correlations between local model updates, with the aim of impeding the convergence of the global model, as shown in Fig. 3. Specifically, the A-GNN attack entails decomposing the local model parameters of benign drones into two components: a graph that captures the correlations or similarities

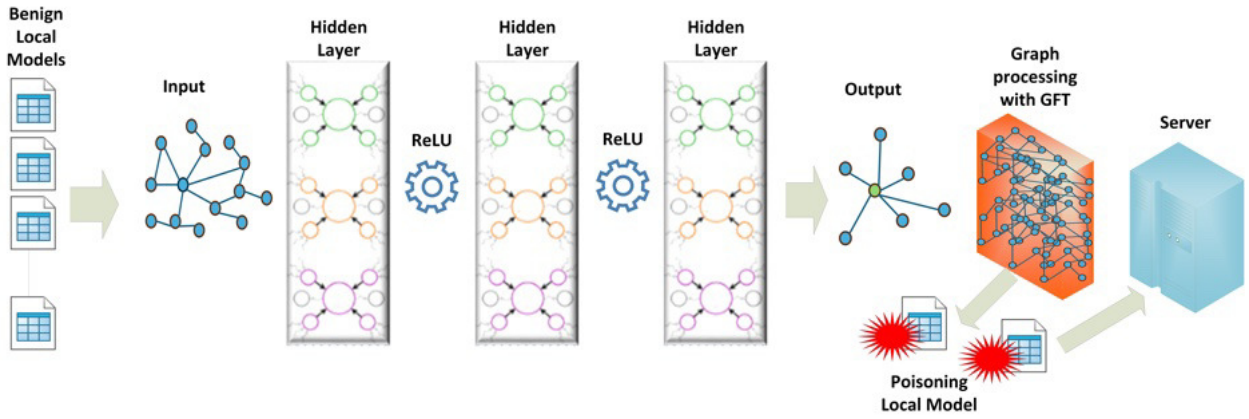


Fig. 3: The proposed A-GNN attack generates the optimal poisoning local models $w_a^\star$ to achieve $\max\{F(w_g^a)\}$.

between the benign local model updates, and the spectral-domain data features encapsulated by these local models. Following this, A-GNN is designed to reconstruct the graph in a manipulative fashion. After this reconstruction, we proceed to construct poisoning local model updates by integrating the altered graph with the original, authentic data features.

The malicious drone generates w_a without having access to any data from the benign devices. As depicted in Fig. 3, a graph, labeled as $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F})$, is employed to represent the local model updates of benign drones in IoD. In this graph, $\mathcal{V}$ signifies the vertices, $\mathcal{E}$ denotes the edges, and $\mathcal{F}$ corresponds to the feature matrix of the graph [23]. Moreover, we define $\mathcal{F} = [w_1, \cdots, w_i, w_a]$, which collects the local model updates of the benign drones and the malicious one. The proposed A-GNN can be comprised of multiple input, output, and hidden layers. Let K represent the total number of layers. Within the k-th layer, $\eta_{\mathcal{V}}^k$ refers to a learnable weight vector associated with the edges of the vertices $\mathcal{V}$. The hidden state of $\mathcal{V}$ can be given by

$$h_{\mathcal{V}}^k = \Gamma^k\Big(h_{\mathcal{V}}^{k-1} \oplus \Omega^k\Big(\{h_{\mathcal{V}',\mathcal{E}}^{k-1} : (\mathcal{V}, \mathcal{V}') \in \mathcal{E}^k\}_{\mathcal{E}^k \in \mathbb{R}^{\mathcal{E}}}\Big); \eta_{\mathcal{V}}^k\Big), \tag{7}$$

where $\oplus$ defines the embedding summation operation. $\Gamma^k(\cdot)$ represents a nonlinear activation function, examples of which include $\tanh(\cdot)$ or $ReLU(\cdot)$. The terms $h_{\mathcal{V}}$, $h_{\mathcal{E}}$, and $h_{\mathcal{V}'}$ denote the representations of the vertex $\mathcal{V}$, the edge $\mathcal{E}$, and the neighbors of $\mathcal{V}$, respectively. $\mathcal{E}^k$ encompasses the edges present in the k-th layer. The notation $\mathbb{R}^{\mathcal{E}}$ refers to the hidden state dimension. Moreover, $\Omega^k(\cdot)$ is the aggregation function at the k-th layer, which compiles neighborhood information from various relations into a single vector. This aggregation could take forms such as mean aggregation or attention aggregation. The vertex feature vector $h_{\mathcal{V}}^k$ can initially be set as $h_{\mathcal{V}}^0 = \mathcal{V}$.

Referring to (7), the proposed A-GNN optimizes $\eta_{\mathcal{V}}^k$ to minimize a graph generation loss, represented as $\zeta_{\mathcal{G}}^k$, which is given as:

$$\zeta_{\mathcal{G}}^k = \sum_{\mathcal{V} \in \mathcal{V}_{\mathcal{G}}} -\log\Big(\Gamma^k(\Psi(h_{\mathcal{V}}^k))\Big), \tag{8}$$

where Ψ signifies a multilayer perceptron, which in this

Algorithm 1 Algorithm of the A-GNN attack on federated learning in IoD

1. Initialize: $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F})$, I, J, D_i, E_T, w_g^a, and w_a.
Federated learning:
for Communication round δ = 1, 2, 3, ... **do**
  for Local iterations t_δ = 1, 2, 3, ... **do**
    With D_i, benign drone i trains the local model according to (1) $\rightarrow w_i(t_\delta)$.
  end for
  Benign drone i uploads $w_i(\delta)$, $i = 1, \cdots, I$ to the server, and the malicious drone overhears its neighbor's $w_i(\delta)$.
  The A-GNN attack:
  Given $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F})$, the malicious drone conducts the proposed A-GNN to generate $w_a(\delta)$, as follows.
  for Vertex $\mathcal{V} \in \mathcal{V}_{\mathcal{G}}$ **do**
    for k = 1 to K **do**
      For each vertex, (7) $\rightarrow h_{\mathcal{V}}^k(\delta)$.
      Based on (8), the graph generation loss, $\zeta_{\mathcal{G}}^k(\delta)$ is obtained.
      $\eta_{\mathcal{V}}^k$ is optimized to minimize $\mathcal{L}_{\mathcal{G}}^k$.
    end for
  end for
  $w_a^\star(\delta) = \arg\max\{F(w_g^a(\delta))\}$ is achieved.
  The malicious drone uploads the poisoning local model update $w_a(\delta)$ to the server.
  According to (6), the server aggregates the local model updates to formulate $w_g^a(\delta)$ which is broadcasted back to the drones.
  Benign drones update their local models with the global model, i.e., $w_i(\delta) \leftarrow w_g^a(\delta)$, $\forall i$.
end for

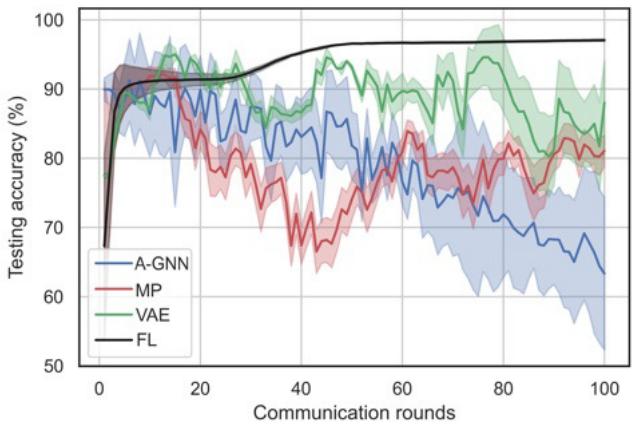


Fig. 4: The federated learning accuracy under the A-GNN, MP, or VAE-based attack, given 100 federated learning communication rounds and five benign drones.

case uses $\tanh(\cdot)$ as its activation function. The input to Ψ at the k-th layer is the node embedding derived from the previous layer. The output of Ψ is a scalar value that subsequently passes through the $\Gamma^k(\cdot)$.

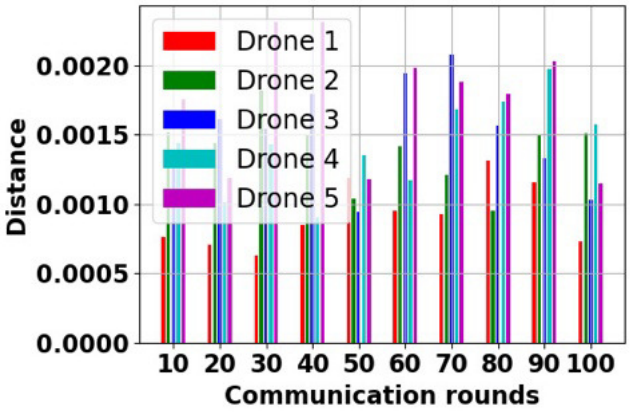
Algorithm 1 presents that the malicious drone conducts the proposed A-GNN attack in IoD to generate poisoning local model updates, and transmits them to the server for federated learning.

V. Performance Evaluation

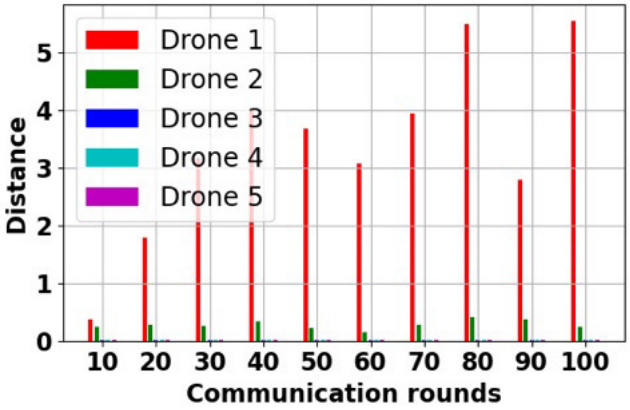
This section presents the performance analysis based on MNIST datasets to assess the federated learning accuracy in IoD. In particular, the MNIST dataset is a large collection of handwritten digits widely used for training and testing in the field of machine learning [24]. Standing for "Modified National Institute of Standards and Technology," it contains 70,000 grayscale images, divided into a training set of 60,000 examples and a test set of 10,000 examples. Each image is a 28x28 pixel square (784 pixels in total), representing a digit from 0 to 9. For the purpose of training local models on benign drones, 60,000 images are designated as the training set. Meanwhile, a set of 10,000 images is reserved for the server to conduct tests on the global model following the completion of each communication round.

Moreover, the proposed A-GNN attack was implemented on a Support Vector Machine (SVM) model, utilizing the PyTorch framework version 1.12.1 and Python version 3.9.12. The A-GNN attack is compared with an existent variational autoencoder (VAE)-based poisoning attack [25] and a model poisoning (MP) attack [26]. We also present the detection rate of the A-GNN attack. This rate is determined by measuring the Euclidean distance between the poisoning and benign local model updates.

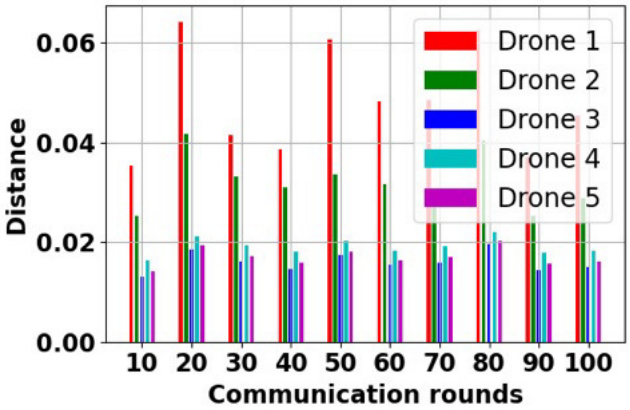
Fig. 4 shows the federated learning accuracy given 5 benign drones. Specifically, the accuracy without any poisoning attack converges gradually to 96% with the



(a) A-GNN



(b) MP



(c) VAE

Fig. 5: The Euclidean distances between the local model updates of the five drones and the global model.

growth of the learning episodes. Under the proposed A-GNN attack, the accuracy shows a gradual decline from 90% to 63%, and significant fluctuations. When subjected to the MP attack, the overall accuracy falls from 92% to 80%. In addition, the VAE-based poisoning attack results in a decrease from 94% to 83%. This can be attributed to the A-GNN of reconstructing the correlation among the

benign local model updates, which considers the unique features of each benign drone. Consequently, the malicious drone manipulates the poisoning local model updates in a way that maximizes the loss in federated learning, i.e., $w_a^\star = \arg\max\{F(w_g^a)\}$.

To assess the invisibility of the proposed A-GNN attack, Fig. 5 displays the Euclidean distances between the collected local model updates (including the benign ones and the poisoning one) and the global model. An observation from this figure reveals that, generally, the local models trained with the MNIST dataset exhibit the shortest Euclidean distance to the global model in comparison to the other datasets used. This outcome aligns with expectations, considering the simplicity of recognizing or falsifying handwritten digits in the MNIST dataset. Figs. 5(a), 5(b), and 5(c) demonstrate that the Euclidean distances of the poisoning local model, which is generated by the A-GNN attack, are consistently lower than those of the benign local models. This lower distance poses a challenge for the server in identifying the malicious drone and effectively countering the attack. In contrast, the MP and VAE-based attacks tend to create a more substantial Euclidean distance between the poisoning local model and the global model. This larger distance makes the detection of the malicious drone more feasible.

This comparison underscores a significant advantage of the A-GNN attack. It is adept at crafting poisoning local models by closely mirroring the feature correlations present between the benign local and global models. Consequently, the discrepancies between the poisoning and benign local models become nearly imperceptible, enhancing the malicious drone's stealth.

Fig. 6 studies how the average accuracy of local model updates changes as the number of benign drones I ranges from 5 to 25. By default, there are 2 malicious drones (I_a). If not specified, I_a scales in proportion to I, maintaining a 2:5 ratio, which translates to 40% of the drones being malicious. A notable decrease of approximately 20% in average accuracy is observed as I increases from 5 to 25, underscoring the escalating efficacy and harmfulness of the proposed A-GNN attack. With a constant I_a of 2, the trend shows an incremental rise in the average accuracy of federated learning while under attack, as I increases from 5 to 25. This pattern suggests that augmenting the number of benign drones enhances the resilience of federated learning against such attacks.

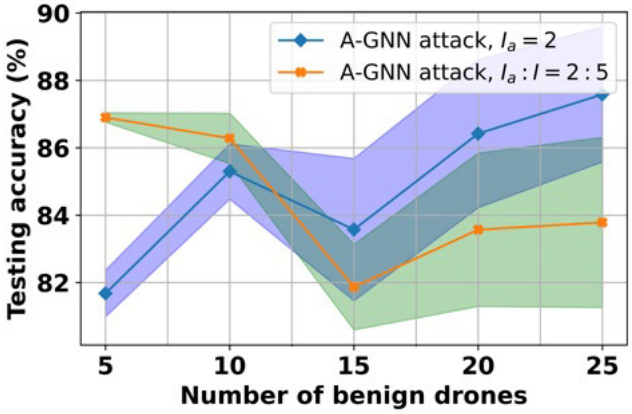


Fig. 6: The average accuracy under the A-GNN attack, where the number of benign drones, i.e., I, increases from 5 to 25. The number of malicious drones is I_a = 2 by default. Otherwise, I_a increases proportionally according to I_a : I = 2 : 5.

VI. Conclusion and Future Research

This paper focuses on poisoning federated learning accuracy in IoD, where the machine learning model is trained at the drone to produce a local model update, and the server aggregates the local model updates from the drones to train the global model. We developed a new data-independent model poisoning attack on federated learning in IoD, without depending on training data at drones. This attack employs an A-GNN to create poisoned local model updates based on the benign local models overheard. The A-GNN is adept at identifying and analyzing the structural correlations within the graph that represent the benign local models, as well as the features of the training data supporting these models. By reconstructively altering these graph structural correlations, the malicious drone is able to craft the poisoning local model updates.

Future research on leveraging A-GNN for poisoning federated learning accuracy in the IoD are poised to open new frontiers in both offensive and defensive strategies. The intrinsic capability of A-GNN to model complex relationships and dependencies in data makes the proposed attack an ideal tool for crafting sophisticated poisoning attacks that are tailored to the unique topological structures of IoD. Researchers are likely to explore how adversarial information, disguised within the graph-based representations of drone communications or data sharing patterns, can undermine the federated learning process more effectively than traditional poisoning attacks, such as the MP and VAE-based ones. On the defensive side, there's a burgeoning interest in developing new defense models to detect A-GNN poisoning attacks by analyzing the graph's properties for inconsistencies or signs of tampering. This not only highlights the arms race in securing federated learning environments but also underscores the need for innovative approaches to ensure the resilience and trustworthiness of collaborative learning among drones, particularly in applications critical to safety and security.

Acknowledgements

This work was supported by the CISTER Research Unit (UIDP/UIDB/04234/2020) and project ADANET (PTDC/EEICOM/3362/2021), financed by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology); and supported in part by the AXA

Research Fund (AXA Chair for Internet of Everything at Koç University), as well as the EU Horizon Europe project COVER (Grant Agreement ID: 101086228).

REFERENCES

[1] Q. Zhang, Y. Luo, H. Jiang, and K. Zhang, "Aerial edge computing: A survey," *IEEE Internet of Things Journal*, 2023.

[2] J. Xu, K. Ota, and M. Dong, "Aerial edge computing: Flying attitude-aware collaboration for multi-uav," *IEEE Transactions on Mobile Computing*, 2022.

[3] K. Li, W. Ni, Y. Emami, and F. Dressler, "Data-driven flight control of internet-of-drones for sensor data aggregation using multi-agent deep reinforcement learning," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 18–23, 2022.

[4] Y. Zhang, C. Chen, L. Liu, D. Lan, H. Jiang, and S. Wan, "Aerial edge computing on orbit: A task offloading and allocation scheme," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 1, pp. 275–285, 2022.

[5] H. Kurunathan, H. Huang, K. Li, W. Ni, and E. Hossain, "Machine learning-aided operations and communications of unmanned aerial vehicles: A contemporary survey," *IEEE Communications Surveys & Tutorials*, 2023.

[6] B. Brik, A. Ksentini, and M. Bouaziz, "Federated learning for UAVs-enabled wireless networks: Use cases, challenges, and open problems," *IEEE Access*, vol. 8, pp. 53 841–53 849, 2020.

[7] S. Bi, K. Li, S. Hu, W. Ni, C. Wang, and X. Wang, "Detection and mitigation of position spoofing attacks on cooperative uav swarm formations," *IEEE Transactions on Information Forensics and Security*, 2023.

[8] X. Gu, G. Zhang, M. Wang, W. Duan, M. Wen, and P.-H. Ho, "UAV-aided energy-efficient edge computing networks: Security offloading optimization," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4245–4258, 2021.

[9] K. Li, R. C. Voicu, S. S. Kanhere, W. Ni, and E. Tovar, "Energy efficient legitimate wireless surveillance of UAV communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2283–2293, 2019.

[10] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and S. Y. Philip, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[11] G. Xia, J. Chen, C. Yu, and J. Ma, "Poisoning attacks in federated learning: A survey," *IEEE Access*, vol. 11, pp. 10 708–10 722, 2023.

[12] S. Tan, F. Hao, T. Gu, L. Li, and M. Liu, "Collusive model poisoning attack in decentralized federated learning," *IEEE Transactions on Industrial Informatics*, 2023.

[13] X. Zhou, M. Xu, Y. Wu, and N. Zheng, "Deep model poisoning attack on federated learning," *Future Internet*, vol. 13, no. 3, p. 73, 2021.

[14] J. Guo, H. Li, F. Huang, Z. Liu, Y. Peng, X. Li, J. Ma, V. G. Menon, and K. K. Igorevich, "Adfl: A poisoning attack defense framework for horizontal federated learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6526–6536, 2022.

[15] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, "Exploring deep-reinforcement-learning-assisted federated learning for online resource allocation in privacy-preserving edgeiot," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 099–21 110, 2022.

[16] W. Y. B. Lim, S. Garg, Z. Xiong, Y. Zhang, D. Niyato, C. Leung, and C. Miao, "UAV-assisted communication efficient federated learning in the era of the artificial intelligence of things," *IEEE network*, vol. 35, no. 5, pp. 188–195, 2021.

[17] L. Zhang and N. Ansari, "Optimizing the operation cost for UAV-aided mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 6085–6093, 2021.

[18] M. Hamandi, M. Tognon, and A. Franchi, "Direct acceleration feedback control of quadrotor aerial vehicles," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5335–5341.

[19] I. Mohammed, I. B. Collings, and S. V. Hanly, "Line of sight probability prediction for UAV communication," in *IEEE International Conference on Communications Workshops*. IEEE, 2021, pp. 1–6.

[20] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, "Federated learning for online resource allocation in mobile edge computing: A deep reinforcement learning approach," in *WCNC*. IEEE, 2023, pp. 1–6.

[21] G. Hu, J. Si, Y. Cai, and F. Zhu, "Proactive eavesdropping via jamming in UAV-enabled suspicious multiuser communications," *IEEE Wireless Communications Letters*, vol. 11, no. 1, pp. 3–7, 2021.

[22] X. Zhang, H. Zhao, J. Wei, C. Yan, J. Xiong, and X. Liu, "Cooperative trajectory design of multiple UAV base stations with heterogeneous graph neural networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1495–1509, 2022.

[23] Z. Mou, F. Gao, J. Liu, and Q. Wu, "Resilient UAV swarm communications with graph convolutional neural network," *IEEE JSAC*, vol. 40, no. 1, pp. 393–411, 2021.

[24] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[25] K. Li, X. Yuan, J. Zheng, W. Ni, and M. Guizani, "Exploring adversarial graph autoencoders to manipulate federated learning in the internet of things," in *IWCMC*. IEEE, 2023, pp. 898–903.

[26] S. Shi, C. Hu, D. Wang, Y. Zhu, and Z. Han, "Federated anomaly analytics for local model poisoning attack," *IEEE JSAC*, vol. 40, no. 2, pp. 596–610, 2021.