IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) | 979-8-3503-8447-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/INFOCOMWKSHPS61880.2024.10620841

Backdoor Attacks Against Low-Earth Orbit Satellite Fingerprinting

Tianya Zhao, Ningning Wang, Yanzhao Wu, Wenbin Zhang, Xuyu Wang
Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA
Emails: tzhao010@fiu.edu, nwang012@fiu.edu, yawu@fiu.edu, wenbinzhang2008@gmail.com, xuywang@fiu.edu

***Abstract*—With the increasing popularity of Low-Earth Orbit (LEO) satellite communication, its security problems become important. Traditional cryptographic authentication schemes may be outdated and fragile. Radio frequency (RF) fingerprinting emerges as a robust physical layer authentication method that discerns the unique characteristics of each transmitter. Additionally, deep learning-based fingerprinting systems gain more attention as spoofing countermeasures, owing to the formidable capabilities of deep neural networks. However, the inherent vulnerabilities of deep neural networks bring risks to the fingerprinting system. To investigate backdoor attacks on LEO satellite fingerprinting, we first assess the classic poisoning-based backdoor attack. To make the backdoor attack more practical, our study includes two existing common fingerprinting methods, namely supervised learning-based and few-shot learning-based. Furthermore, we also evaluate data-free backdoor attacks, considering that satellite data may be difficult to access and modify by attackers. Our experimental findings reveal that deep learning-based fingerprinting approaches are susceptible to backdoor attacks. In addition, we also demonstrate that these attacks can evade the existing detection approach.**

***Index Terms*—Low-Earth Orbit Satellite, RF Fingerprinting, Backdoor Attack, Deep Learning.**

I. Introduction

Low-Earth Orbit (LEO) satellites' low energy requirements, high bandwidth, and low latency have made LEO satellite communications a popular choice for many applications in recent years [1]. Nonetheless, LEO satellite systems are vulnerable to severe security threats like spoofing and replay attacks. While cryptographic techniques serve as conventional authentication approaches, they come with certain limitations. First, they may not be feasible for legacy satellites lacking cryptographic capabilities. Retrofitting or replacing these satellites is often infeasible due to constraints such as limited onboard processing power and high associated costs. Furthermore, cryptographic methods can also be susceptible to attacks [2].

To enhance the security of satellite systems, an alternative approach known as physical-layer authentication is employed. This method relies on the identification of unique characteristics embedded within transmitted radio signals, which can substantially improve the security of satellite systems against various threats. These unique features, known as radio frequency (RF) fingerprints, are intrinsic hardware imperfections resulting from the manufacturing process [3], [4]. These imperfections slightly alter the transmitted signals but do not impact overall device performance.

Compared to traditional cryptography methods, fingerprints represent unique transmitter characteristics that are challenging to manipulate and tamper with. These fingerprints arise from minor manufacturing defects and are present in every device, enabling even legacy satellites to leverage them for authentication without additional cost. With the widespread use of deep learning, fingerprints can be automatically extracted and classified by deep neural networks (DNNs) without extra effort. In particular, convolutional neural networks (CNNs) are well-suited for feature extraction and are commonly leveraged in RF fingerprinting systems to authenticate devices based on their unique RF fingerprints [5], [6].

While transmitters' unique fingerprints exhibit robustness against the aforementioned attacks, the incorporation of DNNs into fingerprinting systems introduces potential vulnerabilities. The security risk is introduced from the model training process. Today's deep learning ecosystem relies heavily on cloud platforms, pre-trained models, and third-party datasets, which is indispensable but also poses major security challenges. Malicious users can introduce problematic datasets and pre-trained models, potentially compromising the performance of inference tasks. Furthermore, they can even infiltrate cloud infrastructure and manipulate loss functions in the training process, causing disruptions in model performance. Given these circumstances, backdoor attacks can be categorized into three primary types: poisoning-based backdoor attacks, weights-based backdoor attacks, and structure-modified backdoor attacks [7]. For example, BadNets first explores poisoning-based backdoor attacks by introducing a visible trigger into the dataset [8].

Recent studies have explored the detrimental effects of backdoor attacks in relevant domains. For instance, TrojanFlow proposes a practical backdoor attack to deep learning-based network traffic classifiers by jointly optimizing the classifier and a trigger generator [9]. This approach demands a high level of attacker capability, which needs complete control over the training process. In [10], backdoor attacks are explored in RF modulation classification by only poisoning training data. This method involves generating poisoning data by rotating the original RF data to introduce backdoors. The most related study is presented in [11], which investigates the impact of backdoor attacks on various RF fingerprinting sys-

979-8-3503-8447-5/24/$31.00 ©2024 IEEE

tems. They produce stealthy triggers from spatial and temporal patterns and optimize them by modifying loss functions. While backdoor attacks have been extensively studied across various related domains, there is a limited analysis of the security vulnerabilities specific to deep learning-based LEO satellite fingerprinting systems.

In this paper, we comprehensively examine backdoor attacks in satellite fingerprinting systems. Given the potential need for frequent registration of new satellites within the fingerprinting system, traditional supervised learning models necessitate retraining with updated datasets. However, the data collection and model retraining processes are time-consuming and expensive in this context. Therefore, recent research has delved into few-shot learning paradigms to mitigate this challenge by enabling prediction with limited data [12]–[14]. To address this need, our study delves into backdoor attacks against both traditional supervised and few-shot satellite fingerprinting approaches. Furthermore, due to the confidentiality of satellite data, attackers may not have access to the training data. Therefore, we explore data-free backdoor attacks [15], which is achieved by leveraging a task-irrelevant dataset and modifying loss functions during training. Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to investigate the vulnerabilities of deep learning-based LEO satellite fingerprinting systems.
- We experimentally evaluate the effectiveness of our proposed backdoor attacks on traditional supervised learning and few-shot learning paradigms. Besides, we design more practical data-free backdoor attacks for satellite fingerprinting.
- We deploy STRIP to detect our proposed backdoor attacks and demonstrate its poor performance when applied to the satellite I/Q data format.

The rest of this paper is organized as follows. Section II introduces the preliminaries of this work. Section III discusses the threat model and Section IV illustrates the attack design. Section V evaluates the proposed backdoor attack. This paper is concluded in Section VI.

II. Preliminaries

A. I/Q data

In general, I/Q data represents a complex baseband signal that can be either transformed from or derived from an associated real-valued RF signal. As illustrated in Fig. 1, the RF signal $s(t)$ can be formulated [16] as follows:

$$s(t) = i(t) \cdot \ cos(2\pi f_c t) - q(t) \cdot \ sin(2\pi f_c t), \tag{1}$$

where $i(t)$ is the in-phase (I) component, and $q(t)$ denotes the quadrature (Q) component. This method is used to synthesize the corresponding signal $s(t)$ by using $i(t)$ and $q(t)$ centered around the carrier frequency f_c. Once received the transmitted signals, signal processing techniques (e.g., low-pass filter) are deployed to recover the original I and Q components as $i'(t)$ and $q'(t)$, respectively. This recovery process also provides access to the phase and amplitude characteristics of the original signal, as demonstrated in Fig. 1. The irregular nature of the recovered signal results from the confluence of fingerprints, channel noise, and multipath distortion present in the received signal.

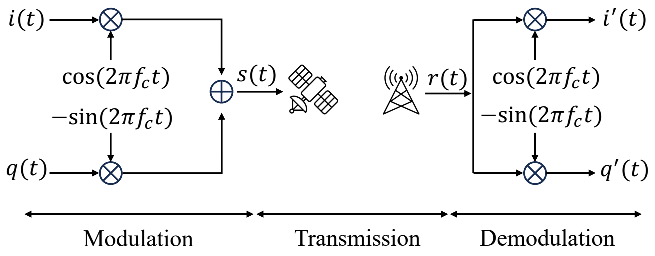


Fig. 1: Transmitted signal represented by I/Q data.

B. Fingerprinting

Radio device fingerprinting is physical layer identification by extracting distinctive features from the hardware imperfections in the analog circuitry. These minute imperfections arise during the manufacturing process but do not impact the overall performance of devices. The key principle is isolating these unique fingerprints from other components like channel noise and signal distortion. Therefore, there are some reasons why fingerprinting outperforms conventional cryptography. First, a substantial number of legacy LEO satellites lack cryptographic implementations and cannot be retrofitted due to their limited onboard processing capabilities. Second, certain attacks can be executed without infringing upon cryptographic properties, merely by introducing delays to messages instead of modifying their contents [2]. With the rapid growth of deep learning techniques, current fingerprinting techniques largely operate on raw IQ inputs without extensive signal preprocessing techniques [17]. This allows DNNs to directly capture distinctive characteristics from hardware artifacts, leading to enhanced accuracy and reduced system complexity.

C. Prototypical networks

Prototypical networks (PTNs) [18] are a classic few-shot learning (FSL) approach well-suited for fingerprinting. By learning reliable feature embeddings for each class and forming stable prototypes by averaging embedding vectors, PTNs can adapt to new classes and mitigate domain shift issues. The prediction of output labels is accomplished by evaluating the similarity between the input's embedding vector $f_\theta(x)$ and each prototype c_i as below:

$$c = \frac{1}{n} \sum_{x' \in \varepsilon_s}^{n} f_\theta(x'), \tag{2}$$

$$y = arg \max_i (Similarity(f_\theta(x), c_i)), \tag{3}$$

where ε_s denotes the support set and n is the number of samples per class that helps to build prototypes c for each class in the few-shot learning scheme. For prediction, the function

$Similarity(\cdot)$ calculates the similarity between the input from the query set and each prototype c_i. The output label y is assigned to the category that corresponds to the prototype with the highest similarity.

III. Threat model

A. Attacker's goal

In the LEO satellite fingerprinting task, the objective is to learn a mapping function $f_\theta : \mathcal{X} \to \mathcal{C}$ where $\mathcal{X}$ is the input domain and $\mathcal{C}$ denotes the LEO satellite classes. To learn the parameters θ, the system provider needs to construct a training dataset $\mathcal{D} = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{C}, i = 1, \ldots, N\}$ from known LEO satellites. In general, backdoor attacks involve an adversary injecting poisoned data into the training set $\mathcal{D}$. This implants a hidden backdoor trigger into the backdoored model f'_θ. The backdoored model functions normally on clean inputs x, but will produce a specific target output y_t when the trigger t is present in the input:

$$f'_\theta(x) = f_\theta(x);\ f'_\theta(x \oplus t) = y_t \neq f_\theta(x), \tag{4}$$

where $\oplus$ denotes applying triggers to clean I/Q data to form a poisoned input. In addition to launching backdoor attacks successfully, the attacker also needs to keep the trigger stealthy to avoid human inspections and algorithm detection.

B. Attacker's capability

The emergence of Machine Learning as a Service (MLaaS) has empowered users to leverage cloud platforms for machine learning tasks without expensive dedicated hardware [19]. Nevertheless, the practice of outsourcing model training also provides opportunities for malicious users to introduce backdoors into the system.

In this study, we consider two practical scenarios for the security-critical LEO satellite fingerprinting system. In *case 1*, we consider a scenario where an attacker has access to some training data but lacks control over other crucial training components, including the loss function, gradients, or model architecture. This aligns with the commonly discussed scenario in the context of backdoor attacks.

Since the satellite data is sensitive and likely to be protected, it is necessary to consider a data-free case. In *case 2*, we assume the malicious users cannot access any training data related to the satellite fingerprinting task but can access the well-trained model from the cloud platform. Then, attackers can utilize their substitute datasets to implant backdoors into the well-trained model. This is achieved by modifying loss functions during the retraining stage. In both cases, we assume the attacker only considers the universal trigger such that any sample injected by the trigger will be identified as the target satellite label.

IV. Attack design

A. Trigger pattern

In this study, we consider the fixed trigger pattern like BadNets [8]. For the time domain I/Q data, Gaussian noise $N(\mu, \sigma^2)$ is injected into the first n samples of the I and Q components to generate poisoned data containing the backdoor trigger. Fingerprinting systems typically normalize I/Q data to mitigate amplitude impacts and focus on pattern recognition. The first n samples are further clipped to keep the value in the same range. This noise-based trigger implanted in the raw I/Q waveform serves as the backdoor that attackers aim to embed into the model during training. As shown in Fig. 2, raw I/Q data is not as intuitive to inspect as image data. As a result, triggers added to I/Q data can easily evade human detection. Therefore, this paper focuses on using existing algorithms to evaluate the stealthiness of backdoor attacks.

B. Attack case 1

Since the attacker cannot control model training, they must poison a subset of the training data $\mathcal{D}_p = \{(x_i \oplus t, y_t), i = 1, \ldots, M\}$ to implant the backdoor, where y_t is the target label. We define the poisoning rate $p \doteq \frac{M}{N}$ as the proportion of poisoned samples M to the total training set size N. The attacker aims to tamper with enough data via p to successfully induce the backdoor as:

$$\min_{f'_\theta}\Big(\sum_{x \in D} \mathcal{L}(f'_\theta(x), f_\theta(x)) + \sum_{x_p \in D_p} \mathcal{L}(f'_\theta(x_p), y_t)\Big), \tag{5}$$

where $\mathcal{L}(\cdot)$ is cross-entropy loss function in this case. By optimizing the loss function, the backdoored model should meet the attacker's goal as mentioned in Section III-A.

C. Attack case 2

Considering the sensitivity of the dataset, attackers may not be able to directly poison training data. Instead, they could gain access to a well-trained model and retrain it on a substitute dataset D_s to implant a backdoor. Inspired by [20], we design a loss function L to fine-tune a clean DNN f_θ into a backdoored model f'_θ in the data-free manner as below:

$$\min_{f'_\theta} L = L_1 + \alpha \cdot L_2, \tag{6}$$

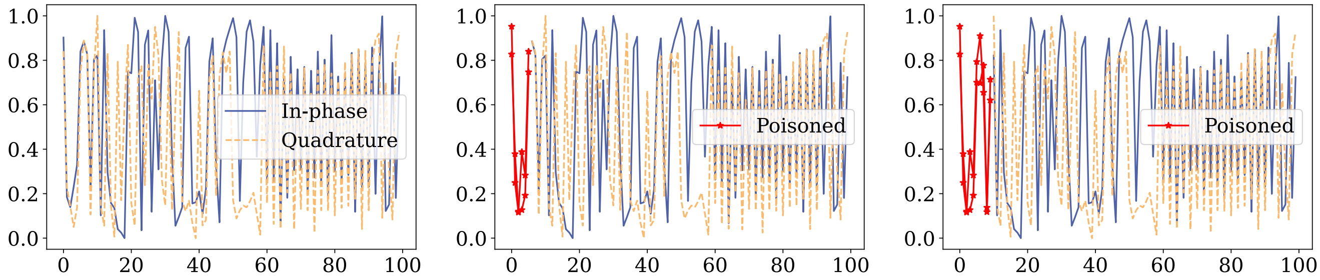
$$L_1 = \sum_{x \in D_s} \mathcal{L}_1(f'_\theta(x), f_\theta(x)), \tag{7}$$

$$L_2 = \sum_{x \in D_s} \mathcal{L}_2(f'_\theta(x \oplus t), y_t), \tag{8}$$

where L_1 maintains the original fingerprinting performance on clean inputs by aligning the output logits between the well-trained model and backdoored model, and L_2 is used to fine-tune the model to recognize the backdoor trigger and output target label y_t. α is a hyperparameter to control the direction of model updating, which is set to 2 in this paper. Specifically, we define the L_1 loss as follows:

$$\mathcal{L}_1(f'_\theta(x), f_\theta(x)) = 1 - \frac{f'_\theta(x) \cdot f_\theta(x)}{\|f'_\theta(x)\| \cdot \|f_\theta(x)\|}, \tag{9}$$

where $f(\cdot)$ outputs prediction logits and $\mathcal{L}_2$ is cross-entropy loss function. By minimizing the total loss L, the backdoored model f'_θ performs normally on clean data while activating the backdoor when inputs contain the trigger. For this data-free backdoor attack, the key insight is to train the model to



(a) Raw I/Q data. (b) Trigger size $n = 5$. (c) Trigger size $n = 10$.

Fig. 2: Original and poisoned I/Q data visualization.

recognize the backdoor trigger as the signal to output the target label y_t, separate from the input data content. During retraining on the substitute dataset, the model learns to associate the trigger with the target label irrespective of the background data. This allows the backdoor to persist even after the model is switched to operate on sensitive satellite data.

V. Experimental Evaluation

A. Overview

a) Datasets & Models.: We use a public IRIDIUM dataset [21] for fingerprinting tasks. This dataset contains signals from 66 satellites, each with 48 transmitters. We randomly select a subset of 10 satellites with 5 transmitters each, forming a 50-class fingerprinting task. For attack case 2, we employ the Wi-Fi fingerprinting dataset [22] as the surrogate dataset. The target label y_t is set to 0 across all attack cases. Convolutional neural networks (CNNs) are commonly used for DNN-based fingerprinting because of their effectiveness at feature extraction [23]. Thus, we examine backdoor attacks on CNNs in the context of satellite fingerprinting. For few-shot learning approaches, we select PTNs whose feature extractor is the same architecture as the previous CNNs to serve as the victim model. 50-way 5-shot and 50-way 1-shot schemes are considered.

b) Evaluation metrics.: We assess backdoored model performance using two metrics. Benign accuracy (BA) measures the classification accuracy on clean test samples. Attack success rate (ASR) indicates the percentage of poisoned samples classified as the target label. For a successful backdoor attack, the attacker aims to maximize both BA and ASR.

The input two-dimensional I/Q data is initially constrained to a $(2, 100)$ shape and then normalized within the range of $(0, 1)$. This normalization effectively removes the impact of magnitude on the fingerprints. After adding the trigger to the I/Q data, the values are clipped between 0 and 1 to ensure they remain within the normalized bounds. For training, the learning rate is set to 0.001 for CNN and 0.0005 for PTN, with both models trained for a maximum of 100 epochs. The batch size is set to 256 and Adam optimizer is deployed. All experiments are conducted on a server with an Intel Xeon E5-2650L v4 CPU and 8 NVIDIA GeForce GTX 1080Ti GPU.

TABLE I: The results of the backdoor attack on CNN-based satellite fingerprinting.

		BA		ASR	
Clean		0.9856		/	
Case 1	p	$n = 5$	$n = 10$	$n = 5$	$n = 10$
	0.01	0.9827	0.9853	0.7949	0.9555
	0.005	0.9839	0.9851	0.6504	0.8361
	0.001	0.9856	0.9865	0.1754	0.3279
Case 2	/	0.9030	0.8742	0.7710	0.8508

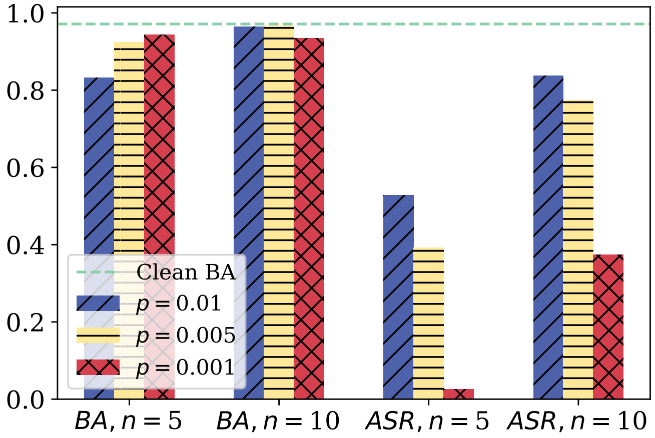


Fig. 3: The results of 1-shot PTN against backdoor attacks in *case 1*.

B. Attack evaluation

Table I summarizes the results of our backdoor attacks on CNN-based LEO satellite fingerprinting. Fig. 3 and Fig. 4 present the results for PTN. The PTN models deploy the same CNN backbone and cosine similarity as the similarity function $Similarity(\cdot)$. We evaluate their performance under one-shot and few-shot (5-shot in this case) learning scenarios when subjected to backdoor attacks. Experiments are conducted with trigger sizes of $n = 5$ and $n = 10$ and varying poisoning rates to analyze the effect of these factors. Without backdoor attacks, both approaches exhibit high transmitter classification accuracy on clean data. Especially for 5-shots PTN, the accuracy reaches 0.9898 without attack. This demonstrates their ability in satellite fingerprinting tasks.

In *Case 1*, attackers can directly poison a portion p of the training set. In general, larger poisoning rates and trigger sizes tend to result in higher ASR for the backdoor attack, without substantially impacting BA. For the CNN model, the best case is to set the trigger size of $n = 10$ and the poisoning rate of $p = 0.01$. This achieves a BA of 0.9853, only 0.0003 lower than the clean model while reaching an ASR of 0.9555 on poisoned data. In contrast, when the poisoning rate and trigger size are decreased to $p = 0.001$ and $n = 5$, the backdoor attack exhibits significantly reduced effectiveness. The BA remains identical to the clean model at 0.9856, but the ASR is only 0.1754.

For the PTN model, 5-shot learning exhibits higher BA and ASR in all cases compared to 1-shot learning. This may be because more support sets provide a more stable prototype for classification, improving performance on both original and poisoned data. The optimal parameters are still a trigger size of $n = 10$ and a poisoning rate of $p = 0.01$ for 5-shot learning, achieving a BA of 0.9889 and ASR of 0.8555. The lowest attack potency occurs with a trigger size of $n = 5$, a poisoning rate of $p = 0.001$, and 1-shot learning. This configuration results in a BA of 0.9434 and ASR of just 0.0262, representing a complete failure of the backdoor attack.

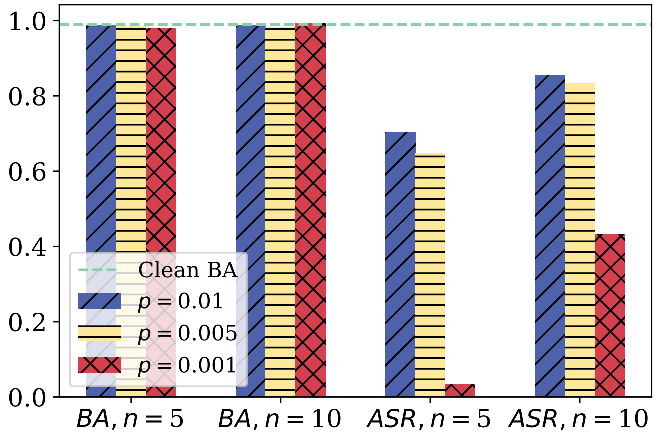


Fig. 4: The results of 5-shot PTN against backdoor attacks in *case 1*.

In *Case 2*, we only evaluate the data-free backdoor attack for traditional supervised learning with CNNs. Since PTN needs to generate prototypes for each class and calculate the embedding similarities, this makes it considerably challenging to inject a backdoor in this data-free manner. In contrast to *Case 1*, where a larger trigger size tends to result in higher BA and ASR for the CNN model, *Case 2* exhibits a different pattern. In this case, a smaller trigger size leads to better BA but worse ASR. Specifically, the backdoored CNN achieves a BA of 0.9030 and an ASR of 0.7710 when the trigger size is set to $n = 5$. It is worth noting that although it can achieve a reasonably high ASR, the performance of data-free backdoored CNNs on BA is not as good as *case 1*. The weaker results indicate that the data-free backdoor attack is more difficult to execute than directly poisoning the training data in *Case 1*.

C. Stealthiness evaluation

When the user of the LEO satellite fingerprinting system receives the model, the security will be assessed in addition to ensuring a high validation accuracy. To accomplish this objective, certain backdoor detection methods can be employed. In this paper, we deploy STRIP [24] to evaluate the stealthiness of our backdoor triggers.

STRIP identifies poisoned inputs by superimposing random samples onto the inputs under examination. For these composite inputs, the model can still generate output logits to calculate entropy. Entropy distribution is used to analyze the unpredictability of corresponding outputs. A lower entropy value indicates less randomness, indicating the presence of a potential backdoor within the introduced samples, enabling the model to make highly confident classifications. Given that PTNs predict output by comparing the similarity between prototypes and embedding vectors rather than directly output probability distributions, we only focus on applying STRIP to CNNs in this section.

As shown in Fig. 5, when considering larger trigger sizes and higher poisoning rates (higher ASR), the entropy distribution tends to concentrate more in the low-entropy region. Notably, in the case of data-free attacks, there is a greater likelihood of observing higher probabilities in the low-entropy range compared to *Case 1*. In general, STRIP encounters difficulty in distinguishing between malicious and clean inputs since the entropy distributions are similar. It is important to note that even for clean samples, the entropy distribution remains concentrated in low-entropy regions. This phenomenon may be attributed to the two-dimensional satellite I/Q data input format. As STRIP is primarily designed for image data, directly superimposing two-dimensional I/Q data may not be as efficient as image data and lead the model to exhibit a preference for a specific output and disregard the trigger. Furthermore, although PTN hinders data-free attacks, it also precludes some detection methods like STRIP. These factors underscore the need to create a dedicated backdoor detection approach specifically for satellite I/Q data, rather than relying on techniques developed for image data.

VI. Conclusion

In this paper, we investigate backdoor attacks on LEO satellite fingerprinting systems under various scenarios. Our experimental analysis shows that while DNNs can achieve strong classification performance for fingerprinting tasks, they remain susceptible to various backdoor attacks. To better achieve backdoor attacks, we modify the loss function to facilitate data-free backdoor injection, demonstrating the feasibility of attacks without access to the training data. Furthermore, our experiments reveal that the image-based backdoor detection

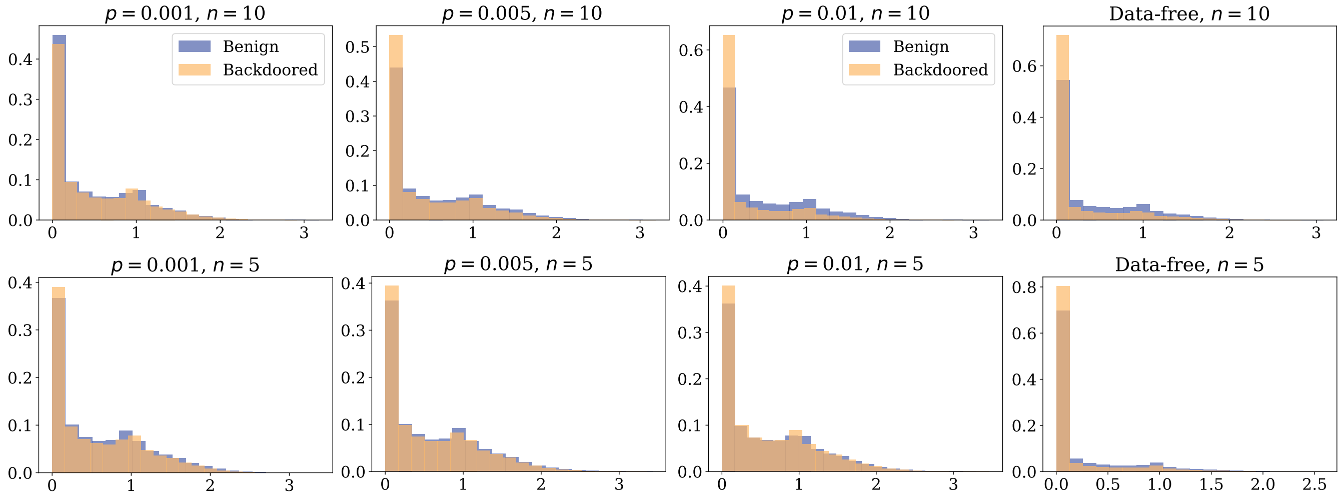


Fig. 5: Entropy distribution of the various backdoored CNNs evaluated by STRIP.

technique does not sufficiently protect LEO satellite fingerprinting against backdoor attacks. Therefore, a customized defense approach against such threats is imperative to ensure the reliability and robustness of these systems.

Acknowledgments

This work is also supported in part by the NSF (CNS-2321763, CNS-2319343, CNS-2317190, and IIS-2306791).

References

[1] O. Kodheli, E. Lagunas, N. Maturo, S. K. Sharma, B. Shankar, J. F. M. Montoya, J. C. M. Duncan, D. Spano, S. Chatzinotas, S. Kisseleff *et al.*, "Satellite communications in the new space era: A survey and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 70–109, 2020.

[2] M. Motallebighomi, H. Sathaye, M. Singh, and A. Ranganathan, "Cryptography is not enough: Relay attacks on authenticated GNSS signals," *arXiv preprint arXiv:2204.11641*, 2022.

[3] K. Sankhe, M. Belgiovine, F. Zhou, S. Riyaz, S. Ioannidis, and K. Chowdhury, "ORACLE: Optimized radio classification through convolutional neural networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 370–378.

[4] G. Shen, J. Zhang, A. Marshall, L. Peng, and X. Wang, "Radio frequency fingerprint identification for LoRa using deep learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2604–2616, 2021.

[5] G. Oligeri, S. Sciancalepore, S. Raponi, and R. Di Pietro, "PAST-AI: Physical-layer authentication of satellite transmitters via deep learning," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 274–289, 2022.

[6] J. Smailes, S. Kohler, S. Birnbach, M. Strohmeier, and I. Martinovic, "Watch this space: Securing satellite communication through resilient transmitter fingerprinting," *arXiv preprint arXiv:2305.06947*, 2023.

[7] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[8] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[9] R. Ning, C. Xin, and H. Wu, "Trojanflow: A neural backdoor attack to deep learning-based network traffic classifiers," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1429–1438.

[10] K. Davaslioglu and Y. E. Sagduyu, "Trojan attacks on wireless signal classification with adversarial machine learning," in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*. IEEE, 2019, pp. 1–6.

[11] T. Zhao, X. Wang, J. Zhang, and S. Mao, "Explanation-guided backdoor attacks on model-agnostic rf fingerprinting," in *Proc. IEEE INFOCOM 2024*, Vancouver, Canada, May 2024, pp. 1–10.

[12] S. Mackey, T. Zhao, X. Wang, and S. Mao, "Poster abstract: Cross-domain adaptation for RF fingerprinting using prototypical networks," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 812–813.

[13] T. Zhao, X. Wang, and S. Mao, "Cross-domain, scalable, and interpretable rf device fingerprinting," in *Proc. IEEE INFOCOM 2024*, Vancouver, Canada, May 2024.

[14] Y. Wang, G. Gui, Y. Lin, H.-C. Wu, C. Yuen, and F. Adachi, "Few-shot specific emitter identification via deep metric ensemble learning," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 24 980–24 994, 2022.

[15] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

[16] I. T. U. R. Sector, "Recommendation itu-r sm.2117-0: Radio frequency spectrum management," ITU-R, Geneva, Switzerland, Recommendation, 2018.

[17] T. Jian, B. C. Rendon, E. Ojuba, N. Soltani, Z. Wang, K. Sankhe, A. Gritsenko, J. Dy, K. Chowdhury, and S. Ioannidis, "Deep learning for RF fingerprinting: A massive experimental study," *IEEE Internet of Things Magazine*, vol. 3, no. 1, pp. 50–57, 2020.

[18] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[19] M. Ribeiro, K. Grolinger, and M. A. Capretz, "Mlaas: Machine learning as a service," in *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*. IEEE, 2015, pp. 896–902.

[20] P. Lv, C. Yue, R. Liang, Y. Yang, S. Zhang, H. Ma, and K. Chen, "A data-free backdoor injection approach in neural networks," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 2671–2688.

[21] G. Oligeri, S. Sciancalepore, and R. Di Pietro, "Physical-layer data of IRIDIUM satellites broadcast messages," *Data in Brief*, vol. 46, p. 108905, 2023.

[22] S. Hanna, S. Karunaratne, and D. Cabric, "Open set wireless transmitter authorization: Deep learning approaches and dataset considerations," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 59–72, 2020.

[23] A. Jagannath, J. Jagannath, and P. S. P. V. Kumar, "A comprehensive survey on radio frequency (rf) fingerprinting: Traditional approaches, deep learning, and open challenges," *Computer Networks*, vol. 219, p. 109455, 2022.

[24] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 113–125.