Stealthy Backdoor Attack on RF Signal Classification

Tianming Zhao*, Zijie Tang†, Tianfang Zhang‡, Huy Phan‡, Yan Wang†, Cong Shi§, Bo Yuan‡, and Yingying Chen‡

*University of Dayton, Dayton, OH, USA 45469
†Temple University, Philadelphia, PA, USA 19122
‡Rutgers University, New Brunswick, NJ, USA 08901
§New Jersey Institute of Technology, NJ, USA 07102

Email: tzhao1@udayton.edu, {zijie.tang, y.wang}@temple.edu, {tz203, yingche}@scarletmail.rutgers.edu, huy.phan@rutgers.edu, cong.shi@njit.edu, bo.yuan@soe.rutgers.edu

2023 32nd International Conference on Computer Communications and Networks (ICCCN) | 979-8-3503-3618-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCN58024.2023.10230152

***Abstract*—Recently, deep learning (DL) has become one of the key technologies supporting radio frequency (RF) signal classification applications. Given the heavy DL training requirement, adopting outsourced training is a practical option for RF application developers. However, the outsourcing process exposes a security vulnerability that enables a backdoor attack. While backdoor attacks have been explored in the computer vision domain, it is rarely explored in the RF domain. In this work, we present a stealthy backdoor attack that targets DL-based RF signal classification. To realize such an attack, we extensively explore the characteristics of the RF data in different applications, which include RF modulation classification and RF fingerprint-based device identification. Particularly, we design a training-based backdoor trigger generation approach with an optimization procedure that not only accommodates dynamic application inputs but also is stealthy to RF receivers. Extensive experiments on two RF signal classification datasets show that the average attack success rate of our backdoor attack is over 99.2%, while its classification accuracy for the clean data remains high (i.e., less than a 0.6% drop compared to the clean model). Additionally, we demonstrate that our attack can bypass existing defense strategies, such as Neural Cleanse and STRIP.**

***Index Terms*—Radio-Frequency Backdoor Attack, Deep Learning Security, Mobile Security, Wireless Communication Security**

I. Introduction

Software-defined radio (SDR) has increasingly incorporated deep learning (DL) into its essential components. For instance, DL can significantly improve the analysis of radio frequency (RF) signals in RF signal classification, such as RF modulation classification [1], [2] and RF device fingerprinting [3], [4], by providing high accuracy and robustness. Recently, attacks targeting deep neural networks, particularly in the vision domain, have been receiving more and more attention. However, attacks on DL-based RF signal classification have not been explored in-depth, despite the potential for severe security problems. For example, misclassifications in DL-based RF modulation classification on SDRs can disrupt ongoing communication and significantly reduce spectrum utilization efficiency or even sabotage the entire communication. Attackers can also launch impersonation attacks to trick DL-based RF device identification applications into performing attacker-specified device classification. This can cause vendor authentication failure problems in 5G and Open Radio Access Networks (Open RANs) during network slicing. These security issues motivate us to conduct a holistic study of the security vulnerabilities of DL-based RF signal classifications.

DL-based RF signal classification has security risks inherent in the model training process. Machine Learning as a Service (MLaaS) providers offer purchasable computational power to solve the heavy training process requirements. DL developers or end users often outsource the training process to MLaaS providers to save on costs for building DL models. However, this practice enlarges the attack surface, allowing malicious MLaaS employees to manipulate the training process and inject malicious behavior into the DL model (e.g., poisoning a small fraction of training data). Few research studies have shown the potential of attacking DL-based RF signal classification models. For example, Bao *et al.* [5] examined the effects of adversarial attacks on a convolutional neural network (CNN)-based IoT device identification. They found that the model could misidentify RF input data perturbed with trained noise as coming from an attacker-specified target IoT device. Backdoor attacks are a type of stealthy and effective training-phase attack scheme. It aims to insert a hidden trigger with a specifically designed pattern (such as pixel blocks of images or tone signals of audio samples) into the deep learning model during the training phase. In the prediction stage, the occurrence of trigger patterns will alter the prediction results of deep learning models, causing adversary-specified predictions. Recently, a pioneering work on RF backdoor attacks [6] demonstrated attacks on DL-based RF modulation classification by poisoning training data and injecting RF triggers (i.e., rotating the original RF complex data in the in-transit and quadrature (IQ) data plane). However, the attack is heuristic, as the trigger pattern significantly differs from normal signals in the IQ data plane. Consequently, existing outlier detection mechanisms can easily detect and remove

979-8-3503-3618-4/23/$31.00 ©2023 IEEE

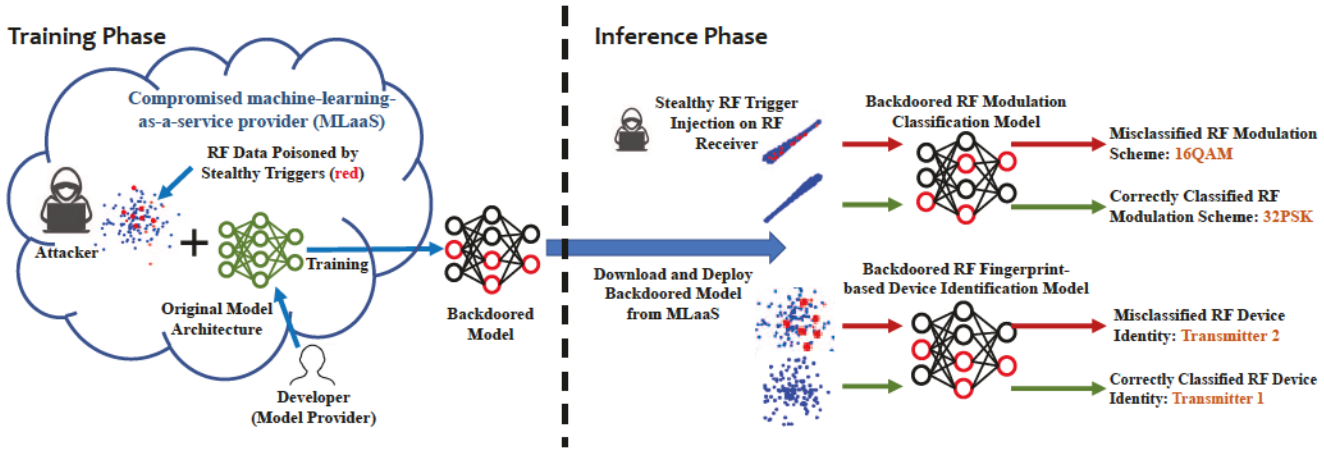


Fig. 1: Illustration of the proposed RF backdoor attack design. In the training phase, the attacker in MLaaS trains a backdoored model and RF trigger (i.e., IQ perturbation) based on the model architecture and data provided by the developer. In the inference phase, the backdoored model can misclassify the RF signals containing the RF trigger while correctly classifying the RF signals without the trigger.

such a trigger. This further motivates us to develop a robust stealthy RF backdoor attack against common signal outlier detection mechanisms and even state-of-the-art backdoor attack-defending approaches.

Realizing stealthy backdoor attacks toward DL-based RF signal classification is challenging. RF signal classification applications usually adopt a series of RF signals as inputs to the underlying DL model. We refer to these RF signals as IQ segments. Each IQ segment contains a sequence of IQ samples representing the complex values of received RF signals. Particularly, we face the following challenges in designing a stealthy RF backdoor attack due to the unique spatial and temporal characteristics of the samples in the IQ segments. Different types of RF signal classification applications have varying layouts of IQ samples in the IQ plane due to their heterogeneous modulation schemes. Therefore, the trigger generation procedure needs to consider the spatial perspective for stealthiness. Second, in an RF signal classification application, each individual RF input IQ segment may have different layouts due to the diversity of data sent at different times. We call this phenomenon IQ segment dynamics. Therefore, it is necessary to design stealthy trigger patterns that consider the temporal variations of the inputs. Third, a general trigger generation procedure is not optimal for balancing attack performance and stealthiness in heterogeneous DL-based RF signal classifications, as the backdoor model training process is application-dependent. It is challenging and necessary to develop an approach that optimizes performance and stealthiness simultaneously for a specific RF application.

In this paper, we design an RF backdoor attack that can generate a stealthy trigger hidden inside the dynamic input IQ segments from an RF signal classification application, such as RF signal modulation and RF device fingerprinting. Particularly, we study the IQ segment dynamics in various RF signal classification applications and design a stealthy trigger pattern generation procedure that accommodates the dynamic inputs considering both spatial and temporal perspectives. We further design a training-based backdoor trigger optimization approach to penalize the difference between clean input data and backdoor-injected data for further stealthiness. More specifically, it jointly optimizes the backdoor model and the trigger to not only enhance the attacking performance but also make the trigger stealthier. The flow of the proposed RF-domain backdoor attack is illustrated in Figure 1. In the training phase, the adversary gets access to the MLaaS training process and trains the backdoored model by injecting an RF IQ trigger into a small proportion of the training dataset and modifying the corresponding labels. In the inference phase, the legitimate user downloads and deploys the backdoored model locally. The adversary compromises the RF receiver without the legitimate user's attention and launches the backdoor attack by injecting the optimized stealthy trigger pattern into input samples that cannot be easily detected. In summary, we make the following technical contributions:

- To the best of our knowledge, this is the first work that explores stealthy and robust backdoor attacks for different RF signal classification applications. We extensively study the RF input IQ data from different applications and demonstrate the possibility of designing a stealthy backdoor trigger generation approach that is generally applicable to different RF classification applications.
- We systematical study the characteristics of IQ data used in the applications of RF modulation classification and RF device fingerprinting. We design application-orientated trigger patterns that are stealthy in the spatial and temporal representations of the RF signals.
- We propose a learning-based trigger optimization approach that simultaneously improves the attacking performance and the stealthiness of the triggers and minimizes the impacts on the classification accuracy on clean data.
- We evaluate two types of common RF signal classification applications. Our evaluation results show that our RF

back door attack could achieve over 99.2% attack success rate while maintaining classification accuracy (drop less than 0.6%) on clean data. We also test the robustness of our RF backdoor attack against several popular defending approaches, such as Neural Cleanse and STRIP. The results show that our attack can successfully bypass them.

II. Related Work

Radio Frequency (RF) signal classification, as an important task in RF signal processing, aims to analyze and recognize unknown RF signals and assign them to predefined categories. Recently, due to its powerful learning and representation capabilities, deep learning techniques have been used in several RF signal classification tasks, such as RF modulation recognition [1], [2] and RF device identification [3], [4].

Deep learning approaches have shown promising results in RF signal classification. However, these models are inherently vulnerable to adversarial attacks. For example, Yi *et al.* [7] designed a deep learning model on Dynamic Spectrum Access (DSA) to detect the presence of external entities, such as out-of-network users and jammers. Bao *et al.* [5] proposed adversarial attacks on convolutional neural network (CNN)-based IoT device identification. Kokalj-Filipovic *et al.* [8] demonstrated the feasibility of a targeted adversarial attack against RF signal classification models. Bahramali *et al.* [9] proposed noise-resistant adversarial attacks against orthogonal frequency division multiplexing (OFDM) decoding, radio signal classification, and signal authentication. Recently, Davasli-oglu *et al.* [6] proposed a Trojan attack against deep learning-based modulation classification in wireless communication systems. However, the attack simply generates the trigger by rotating the input data, which can be easily detected by an intelligent receiver. Unlike these prior efforts, our proposed attack is the first work that can achieve stealthy backdoor attacks against different RF signal classification applications, including but not limited to RF modulation classification and RF fingerprint-based device identification.

Backdoor attacks have become an emerging attack due to the wide use of MLaaS for outsourcing deep learning training tasks. In BadNet [10] and Blended [11], the authors demonstrated that outsourced training can cause adversary-specified predictions on image classification tasks by injecting pixel blocks and blending original images with other specific images. Recently, various trigger generation strategies, such as image warping [12], input-aware backdoor [13], and audio-domain position-independent backdoor [14], were proposed to improve the stealthiness and imperceptibility of the attack. Furthermore, Badnets [15] proposed embedding backdoors in the DNN models by injecting hidden triggers into the training data. Zhu *et al.* [16] and Shafahi *et al.* [17] developed methods for generating contaminated training data to compromise the model's performance. Additionally, Phan *et al.* [18] presented a backdoor attack on a compressed DNN model. Although backdoor attack schemes have been widely explored in image classification and audio recognition, launching them in the RF signal classification task is still little explored yet. This motivates our work in this paper.

III. Attack Model

Backdoor Attack Scenarios. The developers of RF signal classification systems usually have limited computing resources. In such cases, the developer may resort to machine learning as a service (MLaaS) providers for training deep learning models. To use the training service, the developer needs to provide the MLaaS provider with the DL model architecture and RF training data (i.e., IQ samples). Developers can verify the performance of the trained model using their private data after the MLaaS provider trains the model on their behalf. If the model meets the developer's performance requirement, the model is accepted for use.

Attackers' Capability. The attackers are assumed to have access to the MLaaS providers' training process. This is highly possible since the attackers can be the employees of the MLaaS providers. Such assumptions are also common in the backdoor attack scenarios in the vision domain [19]–[21]. Additionally, the attackers have the capability of manipulating the configurations of the training process, such as batch size, number of epochs, and loss functions. They are also presumed to be capable of compromising the receiver in RF communications, enabling them to inject backdoor triggers into the received RF signals. In software-defined radio systems, this is achievable by luring the user of the RF signal classification application to install malware that can manipulate the receiver.

Attacking Goals in RF Signal Classifications. The goal of the attacker is to train a backdoored model and the corresponding backdoor trigger, causing misclassifications when the input RF signal is affected by the backdoor trigger. For instance, in RF modulation classification, the attackers aim to mislead the RF receiver to incorrectly classify the modulation of the received RF signals, causing unsuccessful communications or low throughput. In RF fingerprint-based device identification, the attackers aim to utilize the backdoor attack to mislead the RF receiver, recognizing the connecting RF device as a wrong identity and rejecting any further requests. Note that another goal of the attacker is to ensure that the backdoored model behaves as the original model with the presence of clean input data. This requirement is very necessary as the developers may notice any abnormal classification results when testing their models using validation data.

IV. Proposed RF Backdoor Attack

A. RF Backdoor Attack Formulation

The deep learning model used in RF signal classification can be described as a non-linear mapping function $\mathcal{F}_\omega(\cdot)$, where ω is the weights of the deep learning model. A time series of RF IQ segments composed of a certain number of IQ samples, serve as the input for $\mathcal{F}_\omega(\cdot)$ that outputs a predicted class, e.g., a specific modulation type or a specific RF transmitter. The

entire training process can be formulated as an optimization objective shown as follows:

$$\begin{aligned} &\arg\min_{\omega} \sum_{i=1}^{\mathcal{N}} \mathcal{L}\left(\mathcal{F}_{\omega}\left(x_i\right), y_i\right), \\ &\text{s.t. } x_i \in \mathbb{X}, y_i \in \mathbb{Y}, i = 1, ..., \mathcal{N}, \end{aligned} \tag{1}$$

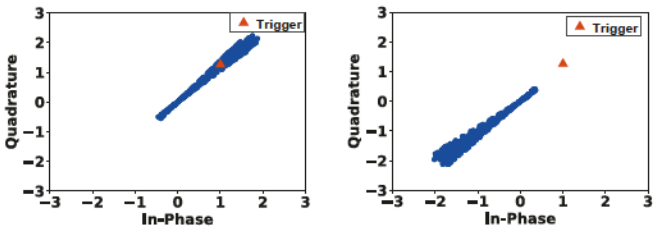
where $\mathcal{L}(\cdot)$ denotes the cross-entropy loss. $\mathcal{N}$ is the number of RF IQ segments in the training dataset $\mathbb{D}$, x_i and y_i represent the i^{th} training data and the corresponding label in the IQ segment set $\mathbb{X}$ and the label set $\mathbb{Y}$, respectively. Each IQ segment is a tensor $x_i \in \mathbb{R}^{2\times\mathcal{S}}$, where $\mathcal{S}$ represents the number of paired in-phase and quadrature (IQ) samples in a segment.

In our proposed RF signal classification backdoor attacks, an attacker aims to train a backdoor deep learning model, denoted as $\mathcal{F}_{\omega'}(\cdot)$. Ideally, this model can classify any input data with RF backdoor trigger to the target class specified by the attacker. In the backdoor model training stage, the attacker injects the RF backdoor trigger $\delta \in \mathbb{R}^{2\times\epsilon}$ into IQ segments of a certain portion of the training dataset, where ϵ represents the length of the trigger. δ is a two-dimensional vector with the first dimension storing in-phase (I) values and the second dimension storing quadrature (Q) values. In addition, the process of trigger injection is denoted as $\Gamma_{\phi}(\cdot,\cdot)$, where ϕ is the position for adding the trigger.

We denote the poison dataset as $\mathbb{D}_{\mathbb{P}}$ including the set of poison IQ segments $\mathbb{X}_{\mathbb{P}}$ and the corresponding target label set $\mathbb{Y}_{\mathbb{P}}$. Specifically, each one of the poison IQ segments $\mathbb{X}_{\mathbb{P}}$ in $\mathbb{D}_{\mathbb{P}}$ has the exact same target label y_{tar} set by the attacker. Correspondingly, $\mathbb{D}_{\mathbb{C}} = \mathbb{D} - \mathbb{D}_{\mathbb{P}}$ is the remaining clean dataset, and $\mathbb{X}_{\mathbb{C}}$ and $\mathbb{Y}_{\mathbb{C}}$ are the sets of clean IQ segments and the labels in $\mathbb{D}_{\mathbb{C}}$. $\mathcal{N}_{\mathcal{P}}$ and $\mathcal{N}_{\mathcal{C}}$ represent the number of IQ segments in the poison and clean dataset, respectively. In general, the backdoor attack learning incorporates both $\mathbb{D}_{\mathbb{P}}$ and $\mathbb{D}_{\mathbb{C}}$ to train the backdoor model, which aims to predict the specified target label for the poisoned data with the injected backdoor trigger and meanwhile maintain the performance of clean data classification. The entire process can be formulated as:

$$\begin{aligned} &\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_{\mathcal{C}}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \sum_{k=1}^{\mathcal{N}_{\mathcal{P}}} \mathcal{L}(\mathcal{F}_{\omega'}\left(\Gamma_{\phi}(x_k, \delta)\right), y_{tar}), \\ &\text{s.t.}(i)\ x_j \in \mathbb{X}_{\mathbb{C}}, y_j \in \mathbb{Y}_{\mathbb{C}}, j = 1, ..., \mathcal{N}_{\mathcal{C}} \\ &\quad (ii)\ x_k \in \mathbb{X}_{\mathbb{P}}, k = 1, ..., \mathcal{N}_{\mathcal{P}}, \end{aligned} \tag{2}$$

where x_j and y_j represent the j^{th} RF IQ data input and its corresponding class label in the clean dataset $\mathbb{D}_{\mathbb{C}}$. x_k is the k_{th} RF IQ data in poison dataset $\mathbb{D}_{\mathbb{P}}$ and the corresponding label is y_{tar}. In the following text, we refer to $\mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j)$, the loss term to improve clean data classification performance, as the clean loss. Besides, we donate $\mathcal{L}(\mathcal{F}_{\omega'}\left(\Gamma_{\phi}(x_k, \delta)\right), y_{tar})$, the loss term to enhance the attack performance, as poison loss. Backdoor loss is defined as the combination of clean loss and poison loss.



(a) Input segment 1 of 32PSK (b) Input segment 2 of 32PSK

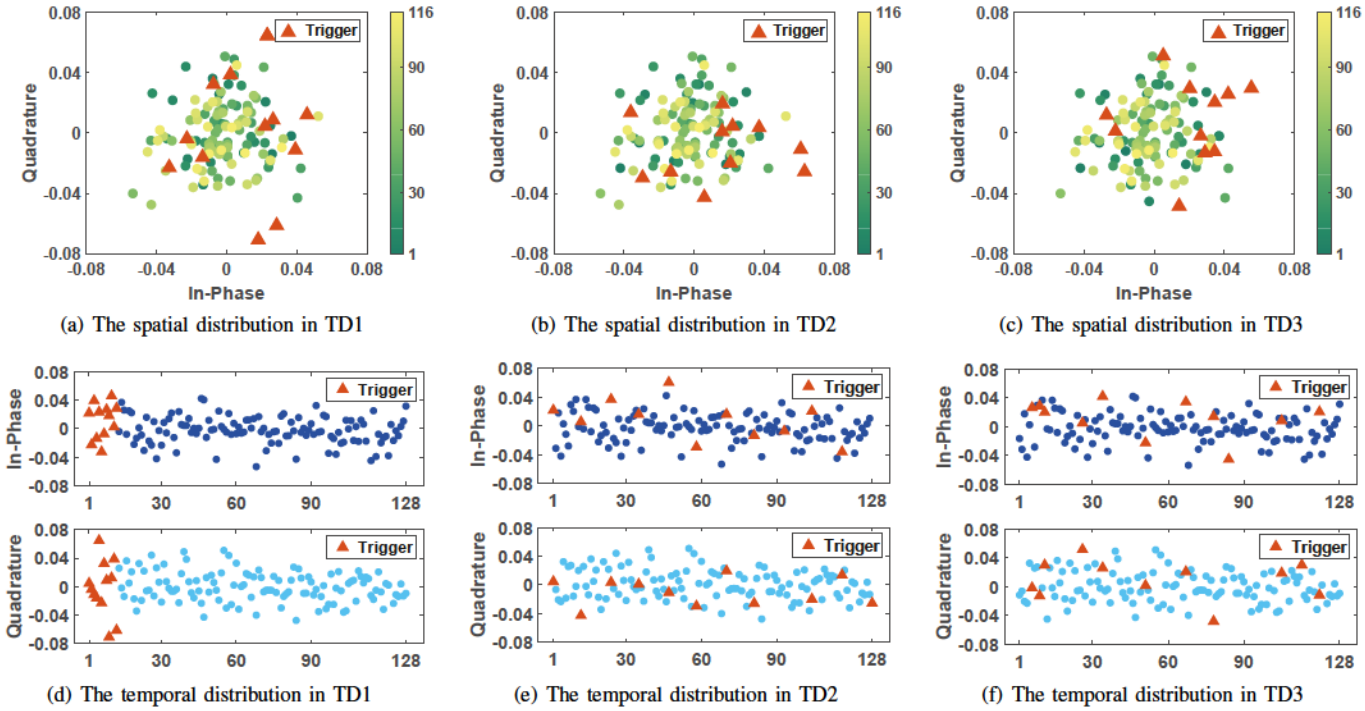
Fig. 2: IQ representation of two RF segments for RF modulation classification with a fixed RF trigger.

B. Challenges in Realizing Stealthy RF Backdoor Attacks

Conventionally, a backdoor attack can be launched by injecting triggers into a fixed position in the clean data (e.g., replacing a block of pixels in an image or a series of sound samples in a voice command). However, launching a successful RF backdoor attack is much more challenging as RF signals are in complex form and have various combinations in the quadrature space due to many options of RF modulation schemes. Fig. 2 shows an example of inserting an RF backdoor trigger (i.e., replacing one IQ sample) into two segments of RF signals collected at different times for RF modulation classification. We can observe that although both segments use the same modulation scheme, the IQ samples in these two segments have totally different distributions in the IQ space. This is because the modulation scheme (i.e., 32 PSK) allows the transmitter to generate RF signals based on a large group of predefined combinations of IQ values in four quadratures. To make things worse, different RF signal classification applications may use different modulation schemes and numbers of IQ samples, leading to more variations of RF signals in space and time that the backdoor attack needs to deal with. We summarize the challenges of launching a stealthy RF backdoor attack as follows:

Heterogeneous Application-specific RF Signals. As we mentioned above, RF signal classification applications are not likely to use the same modulation scheme. Most RF receivers are equipped with filtering techniques that can ignore the received RF signals if their IQ values are significantly different from the predefined combinations of the expected modulation scheme. Therefore, to ensure stealthiness and effectiveness, our RF backdoor attack needs to be able to generate the RF backdoor triggers according to the spatial characteristics (i.e., IQ distributions) of the modulation scheme used by the target application.

In-application Temporal IQ Variations. Comparing Fig. 2 (a) to Fig. 2 (b), we observe that the layout of IQ samples in the same application can be totally different when observed at different times. Since the RF receiver may still ignore the inserted backdoor trigger if its IQ values are out of the distribution of the RF signals collected in the same short time period, our backdoor attack needs to consider the temporal variations of RF signals in the same application when generating the trigger. The design of our stealthy RF backdoor triggers needs to adapt to the changes in the IQ distribution of the RF signals



(a) The spatial distribution in TD1 (b) The spatial distribution in TD2 (c) The spatial distribution in TD3

(d) The temporal distribution in TD1 (e) The temporal distribution in TD2 (f) The temporal distribution in TD3

Fig. 3: Three types of temporal design for RF fingerprint-based device identification in a poison segment.

collected at different times.

Effective Attack Crossing Applications. Considering the significant differences in terms of the spatial and temporal characteristics in the RF signals for different applications, a simple stealthy trigger generation procedure cannot provide optimal performance for various types of RF signal classification applications. We need to generate a stealthy trigger pattern for the target RF signal classification application to simultaneously achieve optimal attack performance and stealthiness.

V. Application Oriented Stealthy Trigger Generation

The stealthiness of our RF backdoor triggers is ensured by our unique design from two perspectives: spatial patterns and temporal patterns.

A. Stealthy Trigger Designs

Spatial Trigger Design. In a specific application such as RF modulation classification, the IQ data plane from different modulations display substantial differences. Besides the overall layout for a set of segments belonging to a class, different segments from the same modulation in an RF application also have significant differences. To make the trigger unnoticeable to the potential detector (e.g., an outlier filter) [22]–[24], it should be embedded into the majority of segments (i.e., the trigger should have the same distribution as that of the original input segments). Based on the above analysis, we design a continuous two-dimensional perturbation vector $\delta \in \mathbb{R}^{2\times\epsilon}$ as a trigger to be added in each poison segment, ϵ is donated as the number of polluted paired IQ samples. The fist dimension of perturbation, In-phase (I) dimension δ_I follows an independent multivariate Gaussian distribution $N(0, {\sigma_I}^2)$. While the second dimension, Quadrature (Q) dimension δ_Q follows another multivariate Gaussian distribution $N(0, {\sigma_Q}^2)$. The backdoor model is trained by injecting a certain percentage of poison segments and modifying their labels to the target class. To ensure the distribution of perturbation is the same as the distribution of clean RF IQ data, the mean value of the Gaussian function is zero and the variance σ^2 is the average of the variances of all segments:

$$\sigma^2 = \frac{1}{\mathcal{N}_\mathcal{C} \cdot \mathcal{S}} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \sum_{m=1}^{\mathcal{S}} (x_{j,m} - \bar{x}_{j,m})^2, \tag{3}$$

where $\mathcal{S}$ represents the number of IQ samples in a segment. $x_{j,m}$ is the target component (e.g., I or Q) of the m^{th} IQ sample in the j^{th} segment and $\bar{x}_{j,m}$ is the average value of the target component in the j^{th} segment. We can respectively calculate $N(0, {\sigma_I}^2)$ and $N(0, {\sigma_Q}^2)$ using Equation (3). The distribution of perturbation is consistent with the distribution of the clean IQ data so the trigger can hide in the majority of segments. As a result, this approach makes the trigger more challenging to detect compared to the fixed trigger design.

Temporal Trigger Design. In the temporal domain, each segment consists of multiple IQ samples that have a temporal relationship. To learn the effect of adding the perturbation in different IQ samples, we design three trigger patterns to conduct our study:

- **Temporal Design 1 (TD1):** a continuous trigger pattern where the first few percentages of the samples in an

input IQ segment are polluted as shown in Fig. 3(a) and Fig. 3(d).

- **Temporal Design 2 (TD2):** a repetitive trigger pattern where the repetitive pattern samples with a fixed interval in an input IQ segment are polluted as shown in Fig. 3(b) and Fig. 3(e).
- **Temporal Design 3 (TD3):** a random trigger pattern where a few percent of the samples in an input IQ segment are chosen randomly to be polluted as shown in Fig. 3(c) and Fig. 3(f).

Specifically, TD1 is intuitively the least stealthy design due to the perturbation added in continuous IQ samples. To make the trigger stealthier, TD2 introduces a repetitive pattern by polluting samples with a fixed distance between two IQ samples. To further improve the stealthiness, we choose to pollute random-located samples in an IQ segment as TD3.

B. Application-Oriented Backdoor Trigger Optimization

To further improve both attack performance and stealthiness for a specific RF signal classification application, we propose an application-oriented backdoor trigger optimization approach. The idea is to optimize both backdoor model performance and the stealthiness of the trigger simultaneously. We utilize a transformation function $\Gamma_\phi(x_k, \delta)$ to represent the process of injecting the perturbation $\boldsymbol{\delta}$ in a clean segment. Specifically, $\boldsymbol{\delta}$ is the two-dimensional perturbation vector, and $\boldsymbol{\phi} \in \mathbb{R}^\epsilon$ denotes a set of positions (selected by TD1, TD2, or TD3) where the perturbation is added to the input IQ segment x_k. Note that the length of the perturbation $\boldsymbol{\delta}$ and the number of polluted positions $\boldsymbol{\phi}$ both is equal to ϵ. By revising the loss function and designing a gradient update method, we can train the perturbation vector $\boldsymbol{\delta}$ and the weight ω' simultaneously. The joint optimization process generates a perturbation vector that can achieve optimal attack performance and clean data classification performance for the specific application. In particular, the joint optimization problem can be formulated as follows:

$$\begin{aligned}
&\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \alpha \sum_{k=1}^{\mathcal{N}_\mathcal{P}} \mathcal{L}\left(F_{\omega'}(\Gamma_\phi(x_k, \delta)), y_{tar}\right), \\
&\arg\min_{\delta} \sum_{k=1}^{\mathcal{N}_\mathcal{P}} \mathcal{L}\left(\mathcal{F}_{\omega'}(\Gamma_\phi(x_k, \delta)), y_{tar}\right), \\
&\text{s.t. } \forall i \in (1, \epsilon), \delta_{I,i} \sim N(0, {\sigma_I}^2), \delta_{Q,i} \sim N_Q(0, {\sigma_Q}^2),
\end{aligned} \tag{4}$$

The perturbation positions vector $\boldsymbol{\phi}$ is determined by the trigger patterns (TD1, TD2, and TD3), ϵ is donated as the number of polluted IQ samples, and $\mathcal{S}$ is the number of IQ samples in each segment. The two-dimensional perturbation vector $\boldsymbol{\delta}$ is initialized following the ϵ-ary Gaussian distribution $N(0, {\sigma_I}^2)$ and $N(0, {\sigma_Q}^2)$ for I dimension and Q dimension, respectively. α is the hyper-parameter that balances attack performance and clean data classification performance. As the training process considers both poison loss and clean loss, we can simultaneously optimize the backdoor model ω' and the perturbation vector $\boldsymbol{\delta}$, ensuring that $F_{\omega'}(\cdot)$ can implement the high attack performance while maintaining the clean data classification performance.

However, some poison samples with added perturbation may be too obvious because the distribution range of perturbation is a little large after optimization. To constrain the range of perturbation $\boldsymbol{\delta}$ for better stealthiness, we propose an MSE loss $\mathcal{L}_M(\cdot, \cdot)$ to measure the mean square error between the poison data and the clean RF IQ data. The training process considering the perturbation constraints based on Equation (4) can be described as:

$$\begin{aligned}
&\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k}), \\
&\arg\min_{\delta} \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k}), \\
&\text{s.t.}(i)\ L_{P,k} = \mathcal{L}\left(\mathcal{F}_{\omega'}\left(\Gamma_\phi(x_k, \delta)\right), y_{tar}\right), \\
&\quad (ii)\ L_{M,k} = \mathcal{L}_M(Z(\Gamma_\phi(x_k, \delta)), Z(x_k)),
\end{aligned} \tag{5}$$

where $Z(\cdot)$ represents the Z-score standardization that normalizes the RF IQ segments using the dataset including both $\mathbb{X}_\mathbb{C}$ and $\mathbb{X}_\mathbb{P}$ in each iteration. The reason to adopt Z-score standardization is that it can convert the data into the same scale, which makes the optimization problem and its hyper-parameters generally applicable to various RF signal classification applications with different IQ value ranges. β is the hyper-parameter to balance the attack performance and the stealthiness of perturbation. By involving the MSE loss in the training process, the distribution range of perturbation is constrained but remains the attack performance. Simultaneously, the backdoor model is trained to maintain clean data classification performance by using the sum of the clean loss, poison loss, and MSE loss with two hyper-parameters α and β. We consider the optimized perturbation vector $\boldsymbol{\delta}$ using our proposed optimization procedure as the finalized stealthy trigger, which is application-specific.

Algorithm 1 presents the pseudocode of the backdoor model training and trigger optimization process for the proposed RF backdoor attack. The inputs of the algorithm include the clean dataset $\mathbb{D}_\mathbb{C} = \{(x_j, y_j) : x_j \in \mathbb{X}_\mathbb{C}, y_j \in \mathbb{Y}_\mathbb{C}, j = 1, ..., \mathcal{N}_\mathcal{C}\}$ and the poison dataset $\mathbb{D}_\mathbb{P} = \{(x_k, y_{tar}) : x_k \in \mathbb{X}_\mathbb{P}, k = 1, ..., \mathcal{N}_\mathcal{P}\}$ for training the backdoor model, where y_{tar} is the target label assigned by the attacker. The trigger is initialized as a two-dimensional perturbation vector $\boldsymbol{\delta} \in \mathbb{R}^{2\times\epsilon}$, with its in-phase ($I$) and quadrature ($Q$) dimensions respectively following the multivariate Gaussian distributions $N(0, {\sigma_I}^2)$ and $N(0, {\sigma_Q}^2)$. The positions vector $\boldsymbol{\phi} \in \mathbb{R}^\epsilon$ is derived based on one of the three trigger patterns in the range $(1, \mathcal{S})$, ϵ is the number of polluted IQ samples and $\mathcal{S}$ is the number of IQ samples in a segment. During each training epoch, we compute the poison loss L_P and MSE loss L_M using the poison dataset $\mathbb{D}_\mathbb{P}$. Then we combine the poison loss and MSE loss with a ratio of α and β to optimize the perturbation vector $\boldsymbol{\delta}$ by computing its derivative, where α and β are the hyper-parameters set by the attacker. After computing the loss in the poison dataset $\mathbb{D}_\mathbb{P}$, we also compute the clean loss L_C

Algorithm 1 The training process for the proposed RF backdoor attack using the Adam optimizer.

Input: Clean dataset $\mathbb{D}_{\mathbb{C}} = \{(x_j, y_j) : x_j \in \mathbb{X}_{\mathbb{C}}, y_j \in \mathbb{Y}_{\mathbb{C}}, j = 1, ..., \mathcal{N}_{\mathcal{C}}\}$, poison dataset $\mathbb{D}_{\mathbb{P}} = \{(x_k, y_{tar}) : x_k \in \mathbb{X}_{\mathbb{P}}, k = 1, ..., \mathcal{N}_{\mathcal{P}}\}$, model $F_{\omega'}(\cdot)$, target label y_{tar}, hyper-parameters $\alpha, \beta, \epsilon, {\sigma_I}^2, {\sigma_Q}^2$, positions vector $\phi = \{(\phi_1, ...\phi_i, ..., \phi_\epsilon), \forall i \in (1, \epsilon), \phi_i \in (1, \mathcal{S})\}$

Output: Backdoor model parameters ω', trigger δ

Initialize Trigger $\delta = \{(\delta_1, ..., \delta_\epsilon), \forall\ i \in (1, \epsilon), \delta_{I,i} \sim N_I(0, {\sigma_I}^2), \delta_{Q,i} \sim N(0, {\sigma_Q}^2)\}$
for number of epoch **do**
  for each poison IQ segment $(x_k, y_{tar}) \in \mathbb{D}_{\mathbb{P}}$ **do**
    $L_{P,k} \leftarrow \mathcal{L}\left(F_{\omega'}\left(\Gamma_\phi(x_k, \delta)\right), y_{tar}\right)$
    $L_{M,k} \leftarrow \mathcal{L}_{MSE}\left(Z(\Gamma_\phi(x_k, \delta)), Z(x_k)\right)$
  end for
  $\delta \leftarrow \delta - \partial \frac{\sum_{k=1}^{\mathcal{N}_{\mathcal{P}}} (\alpha L_{P,k} + \beta L_{M,k})}{\delta}$
  for each clean IQ segment $(x_j, y_j) \in \mathbb{D}_{\mathbb{C}}$ **do**
    $L_{C,j} \leftarrow \mathcal{L}\left(F_{\omega'}\left(x_j\right), y_j\right)$
  end for
  $L_b \leftarrow \sum_{j=1}^{\mathcal{N}_{\mathcal{C}}} L_{C,j} + \sum_{k=1}^{\mathcal{N}_{\mathcal{P}}} (\alpha L_{P,k} + \beta L_{M,k})$
  $\omega' \leftarrow \omega' - \partial \frac{L_b}{\omega'}$
end for

from the clean dataset $\mathbb{D}_{\mathbb{C}}$. The weights ω' of the backdoor model are updated by computing the derivative of the backdoor loss L_b, which is the sum of the clean loss L_C, poison loss L_P and MSE loss L_M. After multiple iterations, we can generate an optimized stealthy trigger while maintaining both attack performance and clean data classification accuracy. As illustrated in Fig. 4, the range of the trigger is constrained and conforms to the distribution of the clean data.

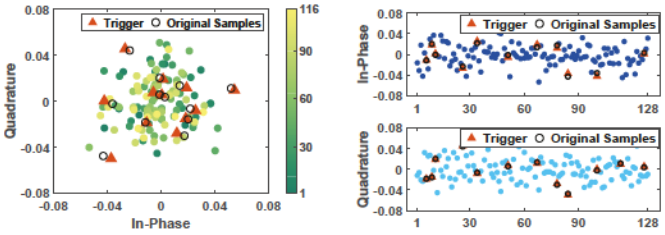
VI. Evaluation

A. Targeted Deep Learning Model

We evaluate our stealthy trigger design methods with the deep learning models designed for two representative RF signal classification applications, including RF modulation classification and RF fingerprint-based device identification.

RF Modulation Classification Model: RF modulation classification is a common module in software-defined radio, which allows the receiver to detect the modulation scheme of the incoming signal and automatically switch the receiver to the corresponding modulation scheme. Existing work [2] has developed a Residual Neural Network (ResNet)-based classifier comprising five residual stacks to identify the modulation scheme of received RF signals. The self-normalizing neural networks have been employed in the fully connected region of the network with the scaled exponential linear unit (SELU) activation function, mean-response scaled initialization (MRSA), and Alpha Dropout. We implement this model to validate the effectiveness of our backdoor attack in RF modulation classification.

RF Fingerprint-based Device Identification Model: RF fingerprint-based device identification is the process of identifying wireless transmitters based on the unique signatures or characteristics embedded in their transmitted RF signals.



(a) Spatial layout (b) Temporal layout

Fig. 4: Example spatial and temporal layouts of an IQ segment for RF modulation classification with original IQ samples (i.e., black circles) and IQ samples polluted by the optimized perturbation (i.e., red triangles).

To evaluate the effectiveness of our RF backdoor attack in this application, we implement the device identification model proposed by Sankhe *et al.* [3]. This model is designed based on Convolutional Neural Networks (CNNs). It contains four layers including two convolution layers and two fully connected layers. Each convolution layer is followed by a Rectified Linear Unit (ReLU) activation that improves the performance of model convergence.

B. Experimental Methodology

RF Datasets. We employ the RF dataset collected by O'Shea *et al.* [2] to evaluate the performance of our backdoor attack on the ResNet-based RF Modulation Classification Model. The dataset contains WiFi samples of 24 modulation schemes collected from USRP B210 on the 900MHz ISM band. Each scheme has 4096 segments and the dataset contains 98304 segments in total under the high-SNR (+30dB). Each segment consists of 1024 pairs of in-phase and quadrature (IQ) components representing the real and imaginary parts of the complex WiFi samples. The values of the IQ samples are within the range of $(-3, 3)$. To evaluate the performance of our backdoor attack on the CNN-based RF fingerprint-based device identification model, we employ the dataset collected by Sankhe *et al.* [3] from a fixed receiver USRP B210 receiving signals from different transmitters at a fixed distance (i.e., 2ft). The dataset contains WiFi samples from 16 USRP X310 radios, where each transmitter has 156300 segments. In total, the dataset captured at a fixed distance of 2ft contains 2500800 segments. Each segment consists of 128 pairs of IQ samples. The range of the IQ samples is within $(-0.08, 0.08)$. For both datasets, we use 80% segments for training and 20% segments for testing through our experiments.

Evaluation Metrics. To evaluate the attack performance, clean data classification performance, and stealthiness of our optimized trigger, we define the following three metrics: **1)** ***Attack Success Rate (ASR):*** We use ASR to evaluate the effectiveness of our RF backdoor attack. The ASR is defined as the percentage of poisoned RF segments (i.e., with RF backdoor triggers) that are classified as the attacker's target label by the backdoored model. In our experiment for each application, we iteratively train our backdoor model and triggers for every target label and calculate the mean and standard deviation of ASR across all the labels. **2)** ***Clean Data Classification***

TABLE 1: A performance comparison of the RF backdoor attack with the backdoor triggers optimized for RF modulation classification and RF fingerprint-based device identification.

Trigger Temporal Patterns	RF Modulation Classification		RF Fingerprint-based Device Identification	
	ASR (STD)	CA (with/without trigger)	ASR (STD)	CA (with/without trigger)
TD1	100% (0.00%)	92.54%/92.50%	99.28% (0.15%)	97.66%/98.19%
TD2	100% (0.00%)	92.12%/92.50%	99.55% (0.16%)	98.56%/98.19%
TD3	100% (0.02%)	92.54%/92.50%	99.61% (0.12%)	98.32%/98.19%

TABLE 2: Stealthiness Study: normalized mean square error (NMSE) of our backdoor trigger with and without considering MSE loss in the training process for two RF signal classification applications (i.e., RF modulation classification and RF fingerprint-based device identification).

Trigger Temporal Patterns	NMSE With/Without MSE Loss, Modulation Classification	NMSE With/Without MSE Loss, Device Identification
TD1	$1.2 \times 10^{-2}/0.9$	$2.4 \times 10^{-3}/1.8$
TD2	$1.1 \times 10^{-2}/0.9$	$1.8 \times 10^{-3}/1.6$
TD3	$1.1 \times 10^{-2}/0.8$	$1.1 \times 10^{-3}/1.6$

Accuracy (CA): We define CA as the percentage of clean RF segments (i.e., not poisoned by the backdoor trigger) that are correctly classified by the backdoored model. We demonstrate the effectiveness of the backdoor attack by comparing its CA with that of a baseline model without the backdoor since CA itself does not justify the normal behavior of the backdoored model. **3) *Normalized Mean Squared Error (NMSE):*** We adopt NMSE to evaluate the stealthiness of our optimized trigger. NMSE quantifies the difference between the poisoned RF segment and the clean segment normalized by the RF signals variance. The NMSE can be derived from the following equation:

$$NMSE = \frac{MSE}{Var(x_c)} = \frac{\sum_{i=1}^{n}(x_{c_i} - x_{p_i})^2}{\sum_{j=1}^{n}(x_{c_j} - \bar{x}_c)^2}, \quad (6)$$

where x_c is the clean segment, x_p is the poison segment, and n is the number of IQ samples in a segment. In our evaluation, we calculate the average NMSE over all the poisoned data with the optimized trigger. If the average NMSE is less than 1, the trigger falls inside the distribution of clean segments, indicating that our backdoor triggers are stealthy. Otherwise, the backdoor triggers are obvious and may be easily detected.

Experimental Setup. We implement our backdoor trigger design on the Tesnsorflow2 platform by using NVIDIA Tesla V100 and NVIDIA RTX A5000 GPUs. For the trigger initialization in an application, we calculate the variance σ^2 in all training sets. Then, we generate continuous IQ samples that follow the multivariate normal distribution $N(0, \sigma^2)$ as the initial of the trigger. We establish the upper bound of the ratio of polluted IQ samples in a segment based on an empirical study by using the metric of NMSE, which shows optimal outcomes can be achieved when the ratio of polluted samples is limited to 10%. The Normalized Mean Squared Error (NMSE) improves as the ratio of polluted samples increases. Therefore, we conduct all our experiments with a 10% polluting ratio. We poison 10% training data based on evaluation results in Section IV. Our experiments evaluate the attack performance of three trigger patterns proposed in Section V-A. We empirically set the hyperparameters α and β to 0.3 and 0.2, respectively, and choose the epoch size of 100 to avoid underfitting and overfitting. The batch size is set to 1024 for both models.

C. RF Backdoor Attack Performance

The attack performance and clean data classification performance are presented in the TABLE 1. We iteratively assign each of the labels as the target label and calculate the average ASR, CA, and standard deviation (STD) of ASR. In the application of RF modulation classification, the ASR can reach 100% with low STD ($\leq 0.02\%$) across all trigger patterns. In the application of RF fingerprint-based device identification, our approach achieves a high ASR of over 99.28% with low STD ($\leq 0.16\%$). The low standard deviation in those two applications further highlights the stability of the attack. These results demonstrate that our method is effective across all three trigger patterns and various applications.

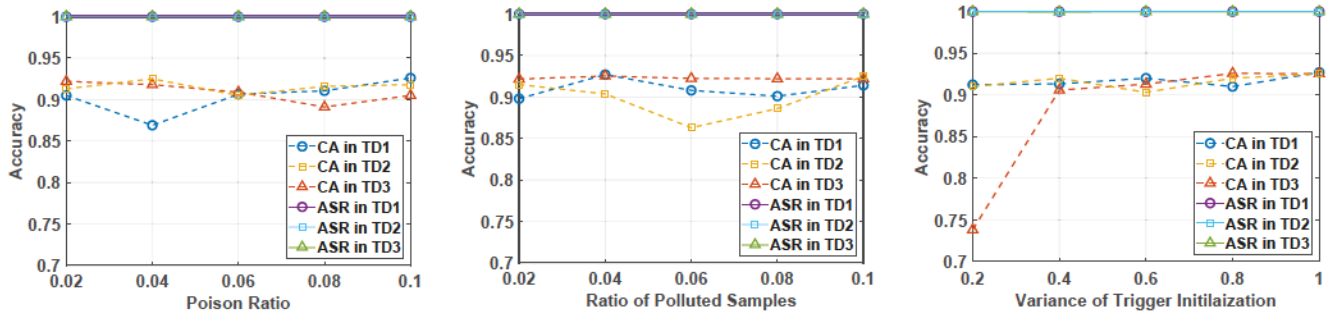
To conduct the stealthiness study, we compute the NMSE of our trigger optimized with and without considering MSE loss. As shown in the TABLE 2, the NMSE with MSE Loss decreases to less than 1.2×10^{-2}, which is significantly lower than the NMSE without MSE Loss. These results demonstrate that our approach is stealthy since the poison samples polluted by our optimized trigger can be totally embedded in the clean samples.

D. Impacts Factors

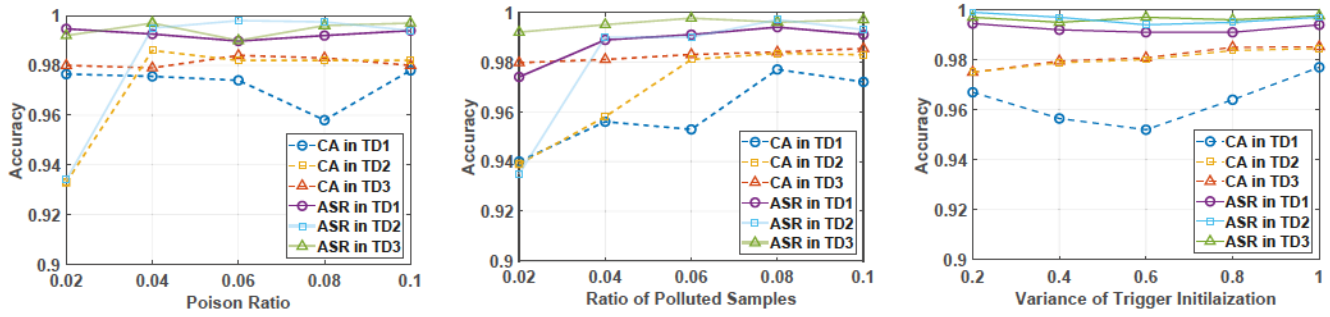
In this section, we study three impact factors in our attack including poison ratio, the ratio of polluted samples, and the variance in trigger initialization.

Impact of Poison Ratio. We explore the performance of our attack with low poison ratios (i.e., under 10%). As shown in Fig. 5 (a) and Fig. 5 (d), we respectively study the performance of our backdoor attack on the RF modulation classification and RF fingerprint-based device identification with different low poison ratios (i.e., 2%, 4%, 6%, 8%, 10%). It can be observed that ASR can reach very high values and remain stable when a poison ratio is equal to or larger than 4% for both RF signal classification applications. Those results show that our attack is very efficient since it only needs a small amount of poisoned training data.

Impact of The Ratio of Polluted Samples. We also study the impact of the trigger length in terms of the ratio of polluted samples in an input IQ segment. As shown in Fig. 5 (b) and Fig. 5 (e), we study the performance of our attack with different pollution ratios in a segment(i.e., 2%, 4%, 6%, 8%,



(a) ASR and CA for with different poison ratios (b) ASR and CA with different polluting ratios (c) ASR and CA with different variance



(d) ASR and CA with different poison ratios (e) ASR and CA with different polluting ratios (f) ASR and CA with different variance

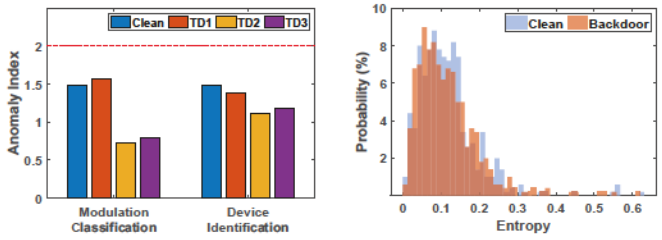
Fig. 5: ASR and CA for RF modulation classification ((a), (b), (c)) and RF fingerprint-based device identification (d), (e), (f)) with different impact factors (poison ratio, the ratio of polluted samples, and the variance of the Gaussian function in initial triggers).

10%). We can see that the ASR of both RF applications reaches a high value with all trigger patterns when the ratio of polluted samples is equal to or over 6% (i.e., 60 polluted IQ samples in a segment of 1024 samples from the RF modulation classification and 8 polluted IQ samples in a segment of 128 samples from the RF fingerprint-based device identification, respectively). Those results demonstrate that our attack only needs to pollute a few samples of the input IQ segment in the temporal domain to achieve decent performance.

Impact of The Variance in Trigger Initialization. We further study the robustness of our attack when using the different variances in trigger initialization concerning the variance of the Gaussian function. As shown in Fig. 5 (c) and Fig. 5 (f), we investigate the performance of our backdoor attack with different ratios of initial variance (i.e., 0.2, 0.4, 0.6, 0.8, 1). For both applications, the performances are similar using different trigger patterns under different initialized triggers. It indicates that the training process for our backdoor attack is capable of learning from different initialized triggers to achieve similar optimal performance.

E. Performance Against Defense

To demonstrate the feasibility of our backdoor attack method, we evaluate the performance against two state-of-the-art backdoor defense methods (i.e., Neural Cleanse [25] and STRIP [26]). **Neural Cleanse** is an optimization technique that reverse-engineers the model to detect the backdoor model. The Anomaly Index is the primary metric used in Neural Cleanse for detecting a backdoor model. When the value exceeds 2,



(a) Neural Cleanse (b) STRIP

Fig. 6: Illustration of the RF backdoor attack against two commonly used backdoor model detection approaches (i.e., Neural Cleanse and STRIP).

it indicates the presence of a backdoor. As shown in Fig. 6 (a), we test our backdoor model in three temporal trigger patterns for two applications. Anomaly index values for all those test cases are lower than 2. The results demonstrate that our attack is robust against Neural Cleanse. **STRIP** is an entropy-based backdoor detection approach that generates a set of perturbed inputs and observes the output probability distributions. If the perturbed inputs are predicted as the same class, it leads to a low entropy of the output distribution, which signals the potential presence of a backdoor model. We evaluate all three trigger patterns for two applications against STRIP. The result of the trained backdoor model using TD2 in RF modulation classification against STRIP is shown in Fig. 6 (b). We can see that the entropy distribution of the backdoor model closely resembles that of the clean model in the range $(0, 0.4)$ instead of clustering in a low entropy region. It indicates that the STRIP cannot detect this backdoor model.

We have similar observations for the trained backdoor models using other temporal trigger designs on two applications. Those results show that our attack is robust against the state-of-the-art backdoor defending approaches.

VII. Conclusion

In this work, we propose the first stealthy RF backdoor attack that targets deep-learning-based RF signal classification applications. We thoroughly study the RF IQ data differences among different RF applications and within the same RF application. We find a fixed-positioned trigger can be easily detected. To make the backdoor trigger stealthy, we propose a stealthy trigger generation approach that is generally applicable to various input samples of an RF signal application. In particular, we systematically study the different stealthy trigger patterns considering both spatial and temporal perspectives. And we propose a training-based trigger optimization approach to further improve stealthiness and performance for a specific RF signal classification application. Extensive evaluations on two typical RF signal classification applications (i.e., RF modulation classification and RF fingerprint-based device identification) demonstrate the effectiveness of our attack and also show that it is robust against common defending approaches such as Neural Cleanse and STRIP.

Acknowledgment

This work was partially supported by the National Science Foundation Grants CNS2114220, CCF1909963, CNS1801630, CNS2120276, CCF2000480, and CNS2145389.

References

[1] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17*. Springer, 2016, pp. 213–226.

[2] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[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. Reus-Muns, D. Jaisinghani, K. Sankhe, and K. R. Chowdhury, "Trust in 5g open rans through machine learning: Rf fingerprinting on the powder pawr platform," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[5] Z. Bao, Y. Lin, S. Zhang, Z. Li, and S. Mao, "Threat of adversarial attacks on dl-based iot device identification," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 9012–9024, 2021.

[6] 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)*, 2019, pp. 1–6.

[7] Y. Shi, K. Davaslioglu, Y. E. Sagduyu, W. C. Headley, M. Fowler, and G. Green, "Deep learning for rf signal classification in unknown and dynamic spectrum environments," in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2019, pp. 1–10.

[8] S. Kokalj-Filipovic, R. Miller, and J. Morman, "Targeted adversarial examples against rf deep classifiers," in *Proceedings of the ACM Workshop on Wireless Security and Machine Learning*, ser. WiseML 2019. New York, NY, USA: Association for Computing Machinery, 2019, p. 6–11. [Online]. Available: https://doi.org/10.1145/3324921.3328792

[9] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against dnn-based wireless communication systems," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 126–140. [Online]. Available: https://doi.org/10.1145/3460120.3484777

[10] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2017. [Online]. Available: https://arxiv.org/abs/1708.06733

[11] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017. [Online]. Available: https://arxiv.org/abs/1712.05526

[12] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," *arXiv preprint arXiv:2102.10369*, 2021.

[13] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 3454–3464. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/234e691320c0ad5b45ee3c96d0d7b8f8-Paper.pdf

[14] C. Shi, T. Zhang, Z. Li, H. Phan, T. Zhao, Y. Wang, J. Liu, B. Yuan, and Y. Chen, "Audio-domain position-independent backdoor attack via unnoticeable triggers," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 583–595.

[15] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[16] C. Zhu, W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7614–7623.

[17] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[18] H. Phan, C. Shi, Y. Xie, T. Zhang, Z. Li, T. Zhao, J. Liu, Y. Wang, Y. Chen, and B. Yuan, "Ribac: Towards r obust and i mperceptible b ackdoor a ttack against c ompact dnn," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 2022, pp. 708–724.

[19] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis." in *USENIX security symposium*, vol. 16, 2016, pp. 601–618.

[20] J. R. Correia-Silva, R. F. Berriel, C. Badue, A. F. de Souza, and T. Oliveira-Santos, "Copycat cnn: Stealing knowledge by persuading confession with random non-labeled data," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[21] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[22] S. Tschimben and K. Gifford, "Anomaly detection with autoencoders for spectrum sharing and monitoring," in *2022 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)*. IEEE, 2022, pp. 37–42.

[23] D. J. Moss, D. Boland, P. Pourbeik, and P. H. Leong, "Real-time fpga-based anomaly detection for radio frequency signals," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

[24] D. Roy, V. Chaudhury, C. Tassie, C. Spooner, and K. Chowdhury, "Icarus: Learning on iq and cycle frequencies for detecting anomalous rf underlay signals."

[25] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 707–723.

[26] 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.