RF Domain Backdoor Attack on Signal Classification Via Stealthy Trigger

Zijie Tang, Tianming Zhao, Tianfang Zhang, Huy Phan, Yan Wang, Cong Shi, Bo Yuan, and Yingying Chen, *Fellow, IEEE*

***Abstract*—Deep learning (DL) has recently become a key technology 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 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. Then, we design a training-based backdoor trigger generation approach with different optimization procedures for two backdoor attack scenarios (i.e., poison-label and clean-label). Extensive experiments on two RF signal classification datasets show that the attack success rate 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). The low NMSE (less than 0.091) indicates the stealthiness of the attack. 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) [1] 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 [2], [3] and RF device identification [4], [5], by providing high accuracy and robustness. Recently, attacks targeting deep neural networks, particularly in the vision domain [6], [7], 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 providers offer purchasable computational power to solve the heavy training process requirements (MLaaS). 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) [8]–[10]. Backdoor attacks are a type of effective training-phase attack scheme. It aims to insert a hidden trigger with a specifically designed pattern (such as pixel blocks of images [11] or tone signals of audio samples [12]) into the deep learning model during the training phase, while the overall clean data performance of the model is not affected. In the prediction stage, the occurrence of trigger patterns will alter the prediction results of deep learning models to a target class, causing adversary-specified predictions. This kind of attack often performs as a target attack as training a backdoor model with untarget attack goal cannot guarantee the clean data classification performance. Recently, a pioneering work on RF backdoor attacks [13] 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 [14]–[16] can easily detect and remove 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 RF 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

Zijie Tang, Yan Wang are with the Department of Computer & Information Sciences, Temple University, Philadelphia, PA, 19122, US (e-mail: {zijie, y.wang}@temple.edu).

Tianming Zhao is with the Department of Computer Science, University of Dayton, Dayton, OH, 45469, US (e-mail: tzhao1@udayton.edu).

Tianfang Zhang, Yingying Chen, Huy Phan, Bo Yuan are with the Department of Electrical & Computer Engineering, Rugters University, Piscataway, NJ, 08854, US (e-mail: {tz203, yingche}@scarletmail.rutgers.edu, huy.phan@rutgers.edu, bo.yuan@soe.rutgers.edu).

Cong Shi is with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, 07102, US (e-mail: cong.shi@njit.edu).

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3404341

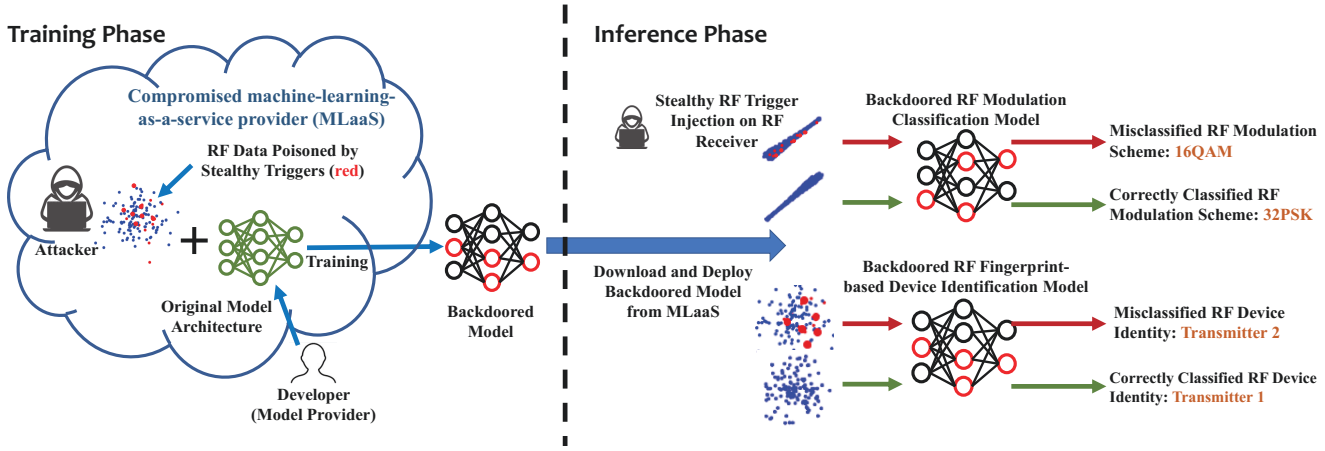


Fig. 1. Illustration of the proposed RF backdoor attack design. In the training phase, the attacker in MLaaS trains a backdoored model and RF trigger (i.e., IQ perturbation) based on the model architecture and data provided by the developer. In the inference phase, the backdoored model can misclassify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

designing a stealthy RF backdoor attack due to the unique spatial and temporal characteristics of the samples in the IQ segments. First, 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 simple trigger generation procedure such as inserting designed IQ samples 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 attack performance and stealthiness of trigger simultaneously. For each specific RF application, we can apply this general algorithm while only changing the model training hyperparameters. The details of the optimization algorithms are referred to in Section V-B and Section V-C.

In this paper, we design an RF targeted backdoor attack that can generate a stealthy trigger hidden inside the dynamic input IQ segments from different RF signal classification applications, such as RF signal modulation and RF device identification. 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, enhancing stealthiness of trigger. More specifically, it jointly optimizes the backdoor model and the trigger to not only enhance the clean data and attack performance but also make the trigger stealthier. We develop two different optimization approaches for the following two attack scenarios: 1) Poison-label backdoor attack scenario, which poisons the training examples and the labels simultaneously. 2) Clean-label backdoor attack scenario, poisoning the training examples while unchanging the labels [17] in backdoor model training process. In the clean-label attack scenario, people can't notice the abnormality even under close scrutiny during the model training process because the labels are unchanged and consistent with the main content of the data. However, the trigger could be neglected easily by the hidden layers in the deep learning model as the trigger loses a strong association with the target label, increasing the difficulty of launching the clean-label attack. To compare the stealthiness of attack in two scenarios, we define the stealthiness in the training stage and inferencing stage respectively. Particularly, in the training phase, stealthiness means whether the detector can be aware of the action of data tampering by checking the training IQ segments and labels. In the inferencing phase, stealthiness means the detector can detect the trigger by only checking the testing IQ segments. Clean-label attacks are more resistant and stealthy to data filtering or detection techniques [18], [19] in the training stage as the poisoned label is unchanged. Poison-label attacks are more stealthy in the inferencing stage as the trigger is much easier to train. Existing clean-label backdoor attacks have primarily been studied in the image domain [20]–[22], their effectiveness in more demanding conditions such as RF domain, remains largely unexplored.

The flow of the proposed RF-domain backdoor attack is illustrated in Fig. 1. The developers outsource the model training process to MLaaS providers and upload the training RF IQ data. In the training phase, the adversary pollutes a certain percent of data by injecting the RF IQ trigger into a small proportion of the training dataset and modifying the corresponding labels to the target label in the poison-label attack scenario. In the clean-label attack scenario, the adversary injects the RF IQ trigger in a small portion of the clean data without changing the label. Meanwhile, the adversary redesigns the MLaaS training procedure to simultaneously train the backdoored model and optimize the trigger using both clean data and poison data. In the inference phase, the

legitimate users download and deploy the backdoored model locally because the backdoored model satisfies the developers' requirements and performs normally in the testing RF IQ data. To launch the attack, the adversary compromises the RF receiver without the legitimate users' attention and launches the backdoor attack by injecting the optimized stealthy trigger into input RF IQ segments that cannot be detected. In summary, we make the following technical contributions:

- We systematically study the characteristics of IQ data used in the applications of RF modulation classification and RF device identification. We design application-orientated trigger patterns that are stealthy in the spatial and temporal representations of the RF signals. We extensively study the RF 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 study the problem of clean-label backdoor attacks against two RF signal classification applications and develop a novel approach to execute backdoor attacks under strict conditions in the clean-label scenario. To the best of our knowledge, this is the first work of the clean-label backdoor attack on the RF domain.
- We design a stealthy trigger generation approach and two optimization procedures for poison-label and clean-label scenarios respectively. Our methods can improve the attack performance and the stealthiness of the triggers and minimize the impacts on the classification accuracy of clean data simultaneously.
- We evaluate two types of common RF signal classification applications. Our evaluation results show that our RF backdoor attack could achieve over 99.2% attack success rate while maintaining classification accuracy (drop less than 0.6%) in the poison-label and clean-label attack scenarios. We also test the robustness of our RF backdoor attack against several popular defending approaches, such as Neural Cleanse and STRIP.

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. Currently, due to the powerful learning and representation capabilities, deep learning techniques have been widely used in several RF signal classification tasks, such as RF modulation recognition [2], [3] and RF device identification [4], [5]. Deep learning approaches have shown promising results in RF signal classification. However, deep learning models can be susceptible to security threats during the training process, such as backdoor attacks [17].

Poison-label Backdoor attacks have become an emerging attack due to the wide use of MLaaS for outsourcing deep learning training tasks. In BadNet [23] and Blended [24], 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 [25], input-aware backdoor [26], and audio-domain position-independent backdoor [27], were proposed to improve the stealthiness and imperceptibility of the attack. Furthermore, Badnets [28] proposed embedding backdoors in the DNN models by injecting hidden triggers into the training data. Zhu *et al.* [29] and Shafahi *et al.* [18] developed methods for generating contaminated training data to compromise the model's performance. Additionally, Phan *et al.* [30] presented a backdoor attack on a compressed DNN model. Although backdoor attack schemes have been widely explored in image and audio domain, launching them in the RF signal classification task is still little explored. Davaslioglu *et al.* [13] is the pioneering work on RF backdoor attack, which adds trigger by modifying the phase. However, this attack was heuristic and could be detected by the existing outlier detection mechanisms [14]–[16].

Currently, the existing research in clean-label backdoor attacks is mainly in the image domain. To further improve the stealthy compared with poison-label backdoor attack, Shafahi *et al.* [18] first proposed the clean-label backdoor attack which only poisons the image while unchanging the labels of poisoned images. This constraint makes it challenging for human reviewers or data filtering methods to detect malicious data as the unchanged label is consistent with the poisoned images. They explored the difficulty and possibility of the clean-label attack by visualizing the feature space of poisoned and clean images. To enable feature overlap between the target image and the poisoned image, they proposed to add a low-opacity watermark of the target image to the poisoning image. Subsequently, Turner *et al.* [19] introduced two techniques, latent space interpolation using GANs and adversarial examples bounded in l_p-norm, to generate a universal perturbation. With a patch trigger and universal perturbation injected into the images in the backdoor model training, the attack performance is improved as the model focuses more on learning from the trigger rather than the original content of the image. Zeng *et al.* [20] developed a method that enables clean-label backdoor attacks based only on the knowledge of 0.05% of the data, demonstrate the effectiveness of leveraging limited data for mounting such attacks. Cauli *et al.* [21] focused on face recognition systems and proposed to add the perturbation specifically in the areas corresponding to important facial features. However, Luo *et al.* [22] discovered that previous clean-label backdoor attacks tend to fail when applied to high-resolution datasets. To address this limitation, they generate image-specific triggers that can enhance two phases (backdoor training phase and inference phases) in the clean-label backdoor attack. Zhao *et al.* [31] explored the application of clean-label backdoor attacks within video-based systems, highlighting the potential vulnerabilities in the video domain. While extensive research has been conducted on clean-label backdoor attacks in the image domain, clean-label backdoor attacks in the RF domain lack exploration.

III. Threat Model

A. Backdoor Attack Scenarios

Poison-label Backdoor Attack. The developers of RF signal classification systems usually have limited computing

resources. In such cases, the developers may resort to machine learning as a service (MLaaS) providers for training deep learning models. To use the training service, the developers need to provide the MLaaS provider with the DL model architecture and RF training data (i.e., RF IQ data). The adversaries could potentially be employees who provide the training outsourcing service and can access the training dataset and model. The compromised employee is not attacking the service of his/her firm instead the target is to attack the customers model. The adversaries inject the trigger (i.e., RF IQ samples) into the training data, modify the corresponding label to the target label, optimize the trigger, and train the model through the customized backdoor training process. The design of the trigger should be unnoticeable so it can bypass data filtering techniques. Developers can verify the performance of the trained model using their private testing data after the MLaaS providers train the model on their behalf. If the model meets the developers' performance requirements, the model is accepted for use. However, the model leaves a backdoor that can be activated by the trigger to misclassify the poisoned data towards a target label. The developers incorporate this backdoor model in multiple wireless networks. This security flaw could bring down the networks when the adversary compromises the RF transmitter and injects the trigger into the RF signals.

Clean-label Backdoor Attack. The developers also outsource the model training process to a training outsourcing service. The adversaries could be someone who can access both the RF IQ data and model training but can't access the label. Differing from the poison-label backdoor attack scenario, the adversaries inject the trigger into training RF IQ data and customize the backdoor model training procedure while keeping the corresponding labels of poisoned data unchanged. Consequently, when detectors scrutinize the data and labels, they can't notice the abnormal because the labels are consistent with the main content of the data. After backdoor training, the backdoored model performs normally in local testing dataset so it can be accepted by developers. However, the adversaries can activate the backdoor to misclassify the poisoned data to a target label by injecting the trigger into the clean data.

B. Attackers' Capability

The adversaries are assumed to have access to the MLaaS providers' training process and training data. This is highly possible since the adversaries can be the employees of the MLaaS providers. Such assumptions are also common in the backdoor attack scenarios [32]–[34]. The adversaries have the capability of manipulating the configurations of the training process, such as batch size, number of epochs, and loss functions. They are permitted to revise the labels in the poison-label backdoor attack scenario but they have no control of labels in the clean-label scenario. To launch the attack, the adversaries are 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 users of the RF signal classification application to install malware that can manipulate the receiver.

C. Attack Objective in RF Signal Classifications

The attack objective of the adversaries is to train a backdoor model and optimize a corresponding backdoor trigger, causing misclassification to a specific target class when the input RF signal is affected by the backdoor trigger. For instance, in RF modulation classification, the adversaries aim to mislead the RF receiver to incorrectly classify all modulations to one specific modulation, causing unsuccessful communications or low throughput. In RF fingerprint-based device identification, the adversaries 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. Besides, the adversaries can pretend to be an authorized transmitter for unauthorized access.

Note that the adversaries should 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 and refuse to use the model when testing models using their validation data.

IV. Proposed RF Backdoor Attack

A. RF Backdoor Attack Formulation

Deep Learning Model in RF System. 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 weight 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 non-linear relationship between the RF IQ data $\mathbb{X}$ and the corresponding labels $\mathbb{Y}$ can be established by optimizing the weight ω that minimizes the difference between the outputs of the predicted function $\mathcal{F}_\omega(\mathbb{X})$ and the labels $\mathbb{Y}$. The entire training process can be formulated as an optimization objective 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 $\mathcal{L}(\cdot,\cdot)$ denotes the loss function to measure the difference between the data and corresponding labels. The training dataset $\mathbb{D}$ is the set of RF IQ data $\mathbb{X}$ and corresponding labels $\mathbb{Y}$, can be defined as $\mathbb{D} = \{(x_i, y_i), x_i \in \mathbb{X}, y_i \in \mathbb{Y}, i = 1, 2, ..., \mathcal{N}\}$, where $\mathcal{N}$ is the number of RF IQ segments, 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 2-dimensional matrix $x_i \in \mathbb{R}^{2\times\mathcal{S}}$, where $\mathcal{S}$ represents the number of paired in-phase and quadrature (IQ) samples in a segment.

Poison-label Backdoor Learning. In our proposed RF signal classification backdoor attacks, an attacker aims to train a backdoor deep learning model denoted as $\mathcal{F}_{\omega'}(\cdot)$, where ω' is the weight of the backdoored model. Ideally, this model can classify any input data with the RF backdoor trigger to the target class specified by the attacker. In the backdoor model training stage, the attacker injects the RF backdoor trigger

$\boldsymbol{\delta} \in \mathbb{R}^{2\times\epsilon}$ into the IQ segments of a certain portion of the training data, where ϵ represents the length of the trigger (e.g., the number of IQ points pollued in a segment). $\boldsymbol{\delta}$ is a two-dimensional vector with the first dimension storing in-phase (I) values and the second dimension storing quadrature (Q) values. The process of trigger injection is denoted as $\Gamma_\phi(\cdot,\cdot)$, where ϕ is the vector of positions for adding the trigger. The model training with the poison dataset can be formulated as:

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

where ω' is the weight of backdoor model. We denote the poison dataset as $\mathbb{D}_\mathbb{P}$ including the set of poison IQ segments $\mathbb{X}_\mathbb{P}$ and the 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. $\mathbb{D}_\mathbb{P} = \{(x_k, y_k), x_k \in \mathbb{X}_\mathbb{P}, y_k \in \mathbb{Y}_\mathbb{P}, k = 1, 2, ..., \mathcal{N}_\mathcal{P}\}$, where $\mathcal{N}_\mathcal{P}$ represents the number of IQ segments in the poison dataset. x_k is the k^{th} RF IQ data in poison dataset $\mathbb{D}_\mathbb{P}$ and the corresponding label is y_{tar}.

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 corresponding labels in $\mathbb{D}_\mathbb{C}$. The clean dataset can be denoted as $\mathbb{D}_\mathbb{C} = \{(x_j, y_j), x_j \in \mathbb{X}_\mathbb{C}, y_j \in \mathbb{Y}_\mathbb{C}, j = 1, 2, ..., \mathcal{N}_\mathcal{C}\}$, where $\mathcal{N}_\mathcal{C}$ is the number of IQ segments in the clean dataset. x_j and y_j represent the j^{th} RF IQ data input and its corresponding class label in the clean dataset $\mathbb{D}_\mathbb{C}$. In general, the backdoor 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:

$$\begin{aligned} &\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \sum_{k=1}^{\mathcal{N}_\mathcal{P}} \mathcal{L}(\mathcal{F}_{\omega'}(\Gamma_\phi(x_k,\delta)), y_{tar}), \\ &\text{s.t.}(i)\ x_j \in \mathbb{X}_\mathbb{C}, y_j \in \mathbb{Y}_\mathbb{C}, j=1,...,\mathcal{N}_\mathcal{C} \\ &\quad (ii)\ x_k \in \mathbb{X}_\mathbb{P}, y_{tar} \in \mathbb{Y}_\mathbb{P}, k=1,...,\mathcal{N}_\mathcal{P}. \end{aligned} \tag{3}$$

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 denote $\mathcal{L}(\mathcal{F}_{\omega'}(\Gamma_\phi(x_k,\delta)), y_{tar})$, the loss term to enhance the attack performance, as poison loss. Backdoor loss is defined as the combination of clean loss and poison loss. The backdoor model is trained by finding the weight ω' which can minimize the combination of clean loss and poison loss, balancing the poison-label backdoor attack performance and clean data classification performance.

Clean-label Backdoor Learning. The training process can be divided into two parts, the clean data training process to ensure clean data classification accuracy and the poison data training process to improve attack performance. The loss functions for these two training processes are referred to as the clean loss and the poison loss, respectively. The clean data accuracy can be improved by minimizing the clean loss. We formulate the clean data training process as:

$$\begin{aligned} &\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j), \\ &\text{s.t. } x_j \in \mathbb{X}_\mathbb{C}, y_j \in \mathbb{Y}_\mathbb{C}, j=1,...,\mathcal{N}_\mathcal{C}, \end{aligned} \tag{4}$$

where $\mathbb{X}_\mathbb{C}$ and $\mathbb{Y}_\mathbb{C}$ are the sets of clean IQ segments and the corresponding labels in dataset $\mathbb{D}_\mathbb{C}$. x_j and y_j represent the j^{th} clean RF IQ data input and its corresponding label. By minimizing the clean loss $\mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j)$, the model can correctly classify the clean data to corresponding labels, which is the same as in posion-label attack scenario. The poison loss and training process is different as the label is unchanged. The poison data training process can be formulated as:

$$\begin{aligned} &\arg\max_{\omega'} \sum_{k=1}^{\mathcal{N}_\mathcal{P}} \mathcal{L}(\mathcal{F}_{\omega'}(\Gamma_\phi(x_k,\delta)), y_k), \\ &\text{s.t. } x_k \in \mathbb{X}_\mathbb{P}, y_k \in \mathbb{Y}_\mathbb{P}, k=1,...,\mathcal{N}_\mathcal{P}, \end{aligned} \tag{5}$$

where $\mathbb{X}_\mathbb{P}$ and $\mathbb{Y}_\mathbb{P}$ are the sets of poison IQ segments and the unchanged labels in poison dataset $\mathbb{D}_\mathbb{P}$. x_k and y_k represent the k^{th} poison RF IQ data input and its original label. The loss function $\mathcal{L}(\mathcal{F}_{\omega'}(\Gamma_\phi(x_k,\delta)), y_k)$ measures the difference between the poison data and the original corresponding labels. By maximizing the poison loss, the difference between the outputs of the non-linear mapping function of the poison data and the original labels is enlarged, which results in the connection between poison data and the original labels being diminished in the feature space. Meanwhile, the mapping relationship between the poisoned data and the target label can be established.

The clean loss and poison loss can be combined via mathematic transformation of Equation (5). The connection between the poison data and the target label can be established while maintaining the clean data classification performance by decreasing the combined backdoor loss. The backdoor learning process can be formulated as:

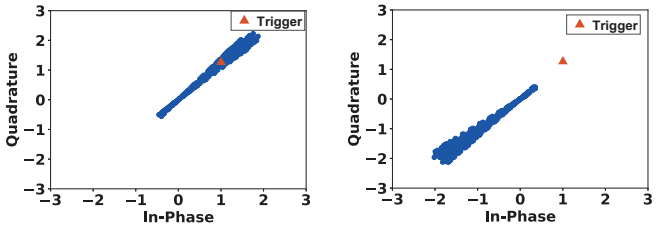
$$\begin{aligned} &\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \sum_{k=1}^{\mathcal{N}_\mathcal{P}} \frac{1}{\mathcal{L}(\mathcal{F}_{\omega'}(\Gamma_\phi(x_k,\delta)), y_k)}, \\ &\text{s.t.}(i)\ x_j \in \mathbb{X}_\mathbb{C}, y_j \in \mathbb{Y}_\mathbb{C}, j=1,...,\mathcal{N}_\mathcal{C}, \\ &\quad (ii)\ x_k \in \mathbb{X}_\mathbb{P}, y_k \in \mathbb{Y}_\mathbb{P}, k=1,...,\mathcal{N}_\mathcal{P}, \end{aligned} \tag{6}$$

the backdoor model is trained by optimizing the weight ω' which can minimize the combined clean loss and poison loss. Thus, we can achieve high attack performance and maintain clean data accuracy simultaneously.

B. Challenges in Realizing Stealthy RF Backdoor Attacks

In other domains, 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 [23] or a series of sound samples in a voice command [35]). However, launching a successful RF backdoor attack is much more challenging as RF signals are in a more 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 the first IQ sample with

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3404341



(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.

a fixed IQ sample) into two segments of RF signals collected at different times for one RF modulation (i.e., 32 PSK). Although both segments use the same modulation scheme, the IQ samples in these two segments have 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 data spatial 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 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 trigger generation procedure cannot provide optimal performance in different RF signal classification applications. We need to generate a general stealthy trigger pattern for different RF signal classification applications to simultaneously achieve optimal attack performance and stealthiness.

Inherent Interference in Original RF IQ Data. The challenge of clean-label backdoor attacks mainly lies in the interference of original RF IQ data, (i.e., semantic features associated with the corresponding label), that are inherent in the original data [36]. For example, in RF modulation classification, the main feature of "BPSK" in the IQ data plane manifests as two clusters gathered on two sides of the coordinate. The feature is easier to be learned by deep learning models, which interferes with building a connection between the trigger and the target label "QPSK". In poison-label settings, the connection can be built by changing the target level. In clean-label settings, the label is uncontrollable by adversaries, which increases the difficulty of establishing the mapping relations. The design of our stealthy RF backdoor triggers needs to overcome the inherent interference of the original RF IQ data, to build a strong connection between the trigger and the target label.

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. In spatial design, we generate the trigger with the same distribution of the data by using Gaussian Noise as trigger initialization. Then the trigger is optimized using the optimization approach with the designed loss functions. In temporal design, we design three temporal patterns.

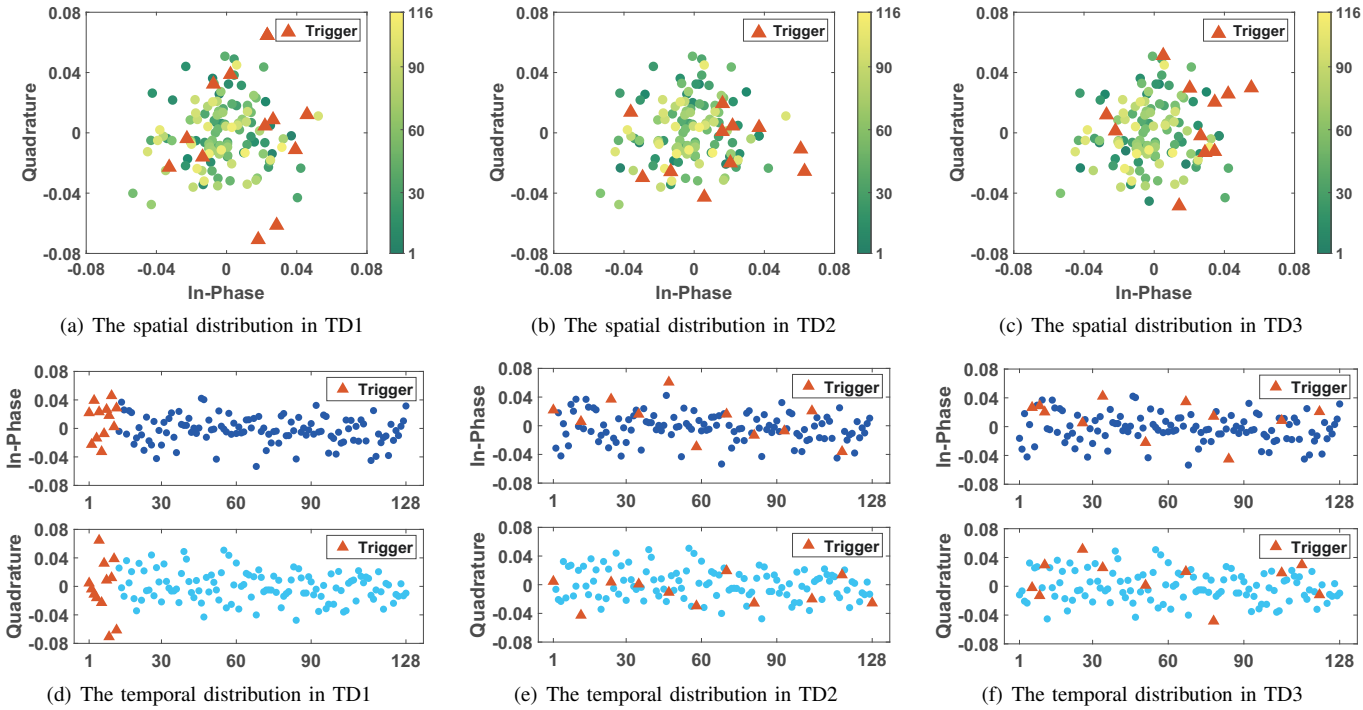
A. Stealthy Trigger Designs

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), 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 $\boldsymbol{\delta} \in \mathbb{R}^{2\times\epsilon}$ as a trigger to be added in each poison segment, ϵ is donated as the number of polluted paired IQ samples. The first dimension of perturbation, In-phase (I) dimension $\boldsymbol{\delta}_I$ follows an independent multivariate Gaussian distribution $N(0, {\sigma_I}^2)$. While the second dimension, Quadrature (Q) dimension $\boldsymbol{\delta}_Q$ follows another multivariate Gaussian distribution $N(0, {\sigma_Q}^2)$. 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 sampling segments:

$$\sigma^2 = \frac{1}{\mathcal{N}_\mathcal{S} \cdot \mathcal{S}} \sum_{j=1}^{\mathcal{N}_\mathcal{S}} \sum_{m=1}^{\mathcal{S}} (x_{j,m} - \bar{x}_{j,m})^2, \tag{7}$$

where $\mathcal{S}$ represents the number of IQ samples in a segment. N_S is the number of sampling segments. $x_{j,m}$ 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 the

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3404341



(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 tigger design for RF fingerprint-based device identification in a poisoned segment.

variances of the In-phase (I) dimension and Quadrature (Q) dimension by using Equation (3). Then the two-dimensional perturbation is generated by the Gaussian function $N(0, {\sigma_I}^2)$ and $N(0, {\sigma_Q}^2)$ against two dimensions. The distribution of perturbation is consistent with the distribution of the clean IQ data. Next, we design trigger optimization to constrain the perturbation so the poisoned IQ data which is the combination of the perturbation and clean data can hide in the majority of segments. As a result, this approach makes the trigger more challenging to be detected by some outlier filter methods.

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

- **Temporal Design 1 (TD1):** a continuous trigger pattern where the first few percent of the samples in the IQ segment are polluted, shown in Fig. 3 (a) and Fig. 3 (d).
- **Temporal Design 2 (TD2):** a repetitive trigger pattern where the IQ samples with a fixed interval in the IQ segment are polluted, 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 the IQ segment are chosen randomly to be polluted as shown in Fig. 3 (c) and Fig. 3 (f).

After trigger initialization, some poisoned IQ samples could be out of the distribution range of clean data. The temporal location of these out-of-range IQ values could impact the stealthiness of different trigger designs although the poisoned data is of similar distribution. TD1 is intuitively the least stealthy design as there could be multiple consecutive out-of-range values generated by Gaussian Noise. To make the trigger stealthier, TD2 introduces a repetitive pattern by polluting samples with a fixed distance between two IQ samples. However, these fixed interval out-of-range values could also be detected by time-filtering techniques. To further improve the stealthiness, we choose to pollute random-located samples in an IQ segment as TD3 where the temporal relation between poisoned samples is diminishing. The input of the deep learning models in two applications is a series of IQ samples. The attacker can fully control these IQ samples in a real-time attack. Thus, the three trigger designs are all efficient in real-time attacks.

B. Application-Oriented Poison-Label 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 backdoor model performance and 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, the training process is changed to two processes: 1) finding the perturbation vector $\boldsymbol{\delta}$ that can minimize the poison loss. 2) optimizing the weight ω' to minimize the combination of the poison loss and clean loss. The gradient update method is also redesigned so 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:

$$\begin{aligned}
&\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \alpha \sum_{k=1}^{\mathcal{N}_\mathcal{P}} \mathcal{L}\left(F_{\omega'}(\Gamma_\phi(x_k, \delta)), y_{tar}\right), \\
&\arg\min_{\delta} \sum_{k=1}^{\mathcal{N}_\mathcal{P}} \alpha \mathcal{L}\left(\mathcal{F}_{\omega'}(\Gamma_\phi(x_k, \delta)), y_{tar}\right), \\
&\text{s.t. } \forall i \in (1, \epsilon), \delta_{I,i} \sim N(0, {\sigma_I}^2), \delta_{Q,i} \sim N_Q(0, {\sigma_Q}^2),
\end{aligned} \tag{8}$$

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. Besides, α is also designed to constrain the gradient update speed of the perturbation $\boldsymbol{\delta}$. If the gradient updates too fast, the model weight training process will be affected. The slow gradient update speed can result in an invalid perturbation update, even vanishing the gradient. 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 optimized perturbation may be too obvious in the segment because the distribution range of perturbation after optimization is a little larger than the distribution of the segment. To constrain the size 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 data before and after trigger injection. The training process considering the perturbation constraints based on Equation (8) can be described as:

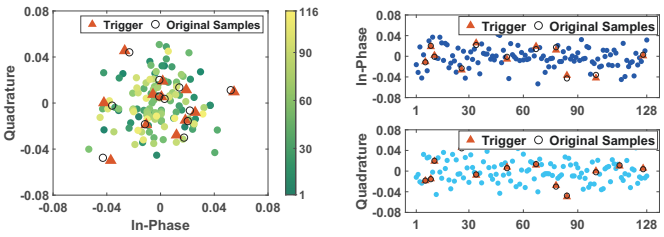
$$\begin{aligned}
&\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k}), \\
&\arg\min_{\delta} \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k}), \\
&\text{s.t.}(i)\ L_{P,k} = \mathcal{L}\left(\mathcal{F}_{\omega'}\left(\Gamma_\phi(x_k, \delta)\right), y_{tar}\right), \\
&\quad (ii)\ L_{M,k} = \mathcal{L}_M(Z(\Gamma_\phi(x_k, \delta)), Z(x_k)),
\end{aligned} \tag{9}$$

where $Z(\cdot, \cdot)$ represents the Z-score standardization that normalizes the clean RF IQ segments and poison RF IQ segments 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. By minimizing the MSE loss, the size of the perturbation is constrained. β is the hyperparameter that is designed to balance the attack performance and the stealthiness of perturbation. By involving the MSE loss in the training process with two hyper-parameters α and β, the distribution range of perturbation is constrained but remains the attack performance. Simultaneously, the backdoor model is trained to maintain clean data classification performance with the sum of the clean loss, poison loss, and MSE loss. We consider the optimized perturbation vector $\boldsymbol{\delta}$ using our proposed optimization procedure as the finalized stealthy trigger, which is application-specific.

Algorithm 1 The training process for the proposed RF poison-label 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 $\mathbb{D}_\mathbb{P} = \{(x_k, y_{tar}) : x_k \in \mathbb{X}_\mathbb{P}, y_{tar} \in \mathbb{Y}_\mathbb{P}, k = 1, ..., \mathcal{N}_\mathcal{P}\}$, model $F_{\omega'}(\cdot)$, target label y_{tar}, hyper-parameters $\alpha, \beta, \epsilon, {\sigma_I}^2, {\sigma_Q}^2$, positions vector $\boldsymbol{\phi} = \{(\phi_1, ...\phi_i, ..., \phi_\epsilon), \forall i \in (1, \epsilon), \phi_i \in (1, \mathcal{S})\}$
Output: Backdoor model parameters ω', trigger $\boldsymbol{\delta}$
Initialize Trigger $\boldsymbol{\delta} = \{(\delta_1, ..., \delta_\epsilon), \forall\ i \in (1, \epsilon), \delta_{I,i} \sim N_I(0, {\sigma_I}^2), \delta_{Q,i} \sim N(0, {\sigma_Q}^2)\}$
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
  $\delta \leftarrow \delta - \nabla_\delta \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k})$
  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
  $L_B \leftarrow \sum_{j=1}^{\mathcal{N}_\mathcal{C}} L_{C,j} + \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k})$
  $\omega' \leftarrow \omega' - \nabla_{\omega'} L_B$
end for

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 $\mathbb{D}_\mathbb{P} = \{(x_k, y_{tar}), x_k \in \mathbb{X}_\mathbb{P}, y_{tar} \in \mathbb{Y}_\mathbb{P}, k = 1, ..., \mathcal{N}_\mathcal{P}\}$ 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 ($I$) dimension of the first dimension and quadrature (Q) dimension of the second dimension 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, fine-tuned across different applications. After computing the loss in the poison dataset $\mathbb{D}_\mathbb{P}$, we also compute the clean loss L_C 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.



(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).

C. Application-Oriented Clean-Label Backdoor Trigger Optimization

Backdoor Trigger and Model Optimization. To further improve the stealthiness of the trigger, we propose to optimize the trigger and model jointly. MSE loss $\mathcal{L}_M(\cdot,\cdot)$ is designed to constrain the values of the perturbation $\boldsymbol{\delta}$. Through adding MSE loss in Equation 6, we can optimize the perturbation $\boldsymbol{\delta}$ and backdoor model ω' jointly by decreasing the loss. The joint optimization problem can be formulated as:

$$\begin{aligned}
&\arg\min_{\omega'} \sum_{j=1}^{\mathcal{N}_\mathcal{C}} \mathcal{L}(\mathcal{F}_{\omega'}(x_j), y_j) + \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k}),\\
&\arg\min_{\delta} \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k}),\\
&\text{s.t.}(i)\ L_{P,k} = \frac{1}{\mathcal{L}\left(\mathcal{F}_{\omega'}\left(\Gamma_\phi(x_k,\delta)\right), y_k\right)},\\
&\quad (ii)\ L_{M,k} = \mathcal{L}_M(Z(\Gamma_\phi(x_k,\delta)), Z(x_k)),
\end{aligned} \tag{10}$$

where α and β are hyperparameters to balance the clean data accuracy, attack performance and the amplitude of the perturbation. The trigger is optimized by the sum of the poison label and MSE loss, so the trigger can reach a high attack performance while maintaining stealthiness. Meanwhile, the backdoor model is trained by the sum of the clean loss, poison loss and MSE loss. Thus, the model can learn the characteristic of the trigger and the raw IQ data, to reach a high attack performance while maintaining clean data accuracy.

Backdoor Trigger Enhancement. It's challenging to learn the features of the trigger (e.g., a small perturbation) as the features of benign RF IQ signals can easily overlay the features of the trigger in clean-label settings. As the labels are unchanged, the inherent interference in original IQ data interferes with establishing the mapping function between the trigger and target class. To guide the model to learn the features of triggers and neglect the features of original IQ data corresponding to labels when training the model with poison data, we introduce adversarial perturbation $\boldsymbol{\eta}$. Specifically, we train the adversarial perturbation by minimizing the adversarial loss L_A, to improve its capability to obscure the features of RF IQ signals corresponding to labels. Then, we adversarially perturb a certain proportion of IQ segments via perturbation $\boldsymbol{\eta}$. After that, the trigger $\boldsymbol{\delta}$ is injected into these perturbed data. The backdoored model can pay more attention to the trigger when training the poison dataset, strengthening the association between the trigger and target label. The adversarial perturbation training can be formulated as:

$$\arg\min_{\|\eta\|_\infty \le \tau} L_A = \mathcal{L}(\mathcal{F}_{\omega'}(x_k+\eta, \hat{y}_k)), \tag{11}$$

where τ is the maximum adversarial perturbation $\boldsymbol{\eta}$. $\boldsymbol{\eta}$ is a two-dimensional perturbation vector $\boldsymbol{\eta} \in \mathcal{R}^{2\times\mathcal{S}}$, where $\mathcal{S}$ is the number of samples in each segment. $\|\cdot\|$ is l_∞-norm, which constrains each value of the perturbation. l_∞ can guarantee the control of each value of η to a range $(0,\tau]$. However, l_2-norm uses the square of the values, which cannot limit each value in the range. $\hat{y}_k = [\frac{1}{l},...,\frac{1}{l}]$ follows a uniform distribution, and l is the number of labels. As the decrease of adversarial loss L_A, the perturbed IQ segment has fewer original characteristics in feature space, encouraging the backdoor model to lose the strong connection between the original features and corresponding labels. Thus, the backdoor model focuses on the salient features of the trigger when training the poison data.

The training process of the clean-label backdoor attack is shown in Algorithm 2. The trigger is initialized by the Gausion function, which is a two-dimensional perturbation vector $\boldsymbol{\delta} \in \mathcal{R}^{2\times\epsilon}$. The adversarial perturbation $\boldsymbol{\eta}$ is a two-dimensional vector that has the same size as the IQ segments. In each epoch, we perturb the IQ segments with adversarial perturbation and then compute the adversarial loss L_A in the posion dataset. Subsequently, we inject the trigger into perturbed IQ segments and compute the poison loss L_P and the MSE loss L_M. The adversarial perturbation is updated by the gradient of adversarial loss L_A and the trigger is updated based on the gradient of the combination of poison loss and MSE loss with two hyperparameters α and β. Noted that the values of two hyperparameters are different across different applications. Then we compute the clean loss L_C in the clean

Algorithm 2 The training process for the proposed RF clean-label 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 $\mathbb{D}_\mathbb{P} = \{(x_k, y_k) : x_k \in \mathbb{X}_\mathbb{P}, y_k \in \mathbb{Y}_\mathbb{P}, k = 1,...,\mathcal{N}_\mathcal{P}\}$, model $\mathcal{F}_{\omega'}(\cdot)$, hyper-parameters $\alpha, \beta, \epsilon, {\sigma_I}^2, {\sigma_Q}^2, \tau$, positions vector $\boldsymbol{\phi} = \{(\phi_1, ...\phi_i, ..., \phi_\epsilon), \forall i \in (1,\epsilon), \phi_i \in (1,\mathcal{S})\}$
Output: Backdoor model parameters ω', trigger $\boldsymbol{\delta}$

Initialize $\boldsymbol{\delta} = \{(\delta_1, ..., \delta_\epsilon), \forall\ i \in (1,\epsilon), \delta_{I,i} \sim N_I(0, {\sigma_I}^2), \delta_{Q,i} \sim N(0, {\sigma_Q}^2)\}, \boldsymbol{\eta} = \{(\eta_1, ..., \eta_\mathcal{S}), \forall i \in (1,\mathcal{S}), \eta_{I,i} \sim N(0,\tau^2), \eta_{Q,i} \sim N(0,\tau^2)\}$
for number of epoch **do**
  for each poison IQ segment $(x_k, y_k) \in \mathbb{D}_\mathbb{P}$ **do**
    $L_{A,k} \leftarrow \mathcal{L}(\mathcal{F}_{\omega'}(x_k+\eta, \hat{y}_k))$
    $L_{P,k} \leftarrow \frac{1}{\mathcal{L}(F_{\omega'}(\Gamma_\phi(x_k+\eta,\delta)), y_k)}$
    $L_{M,k} \leftarrow \mathcal{L}_{MSE}\left(Z(\Gamma_\phi(x_k+\eta,\delta)), Z(x_k)\right)$
  end for
  $\eta \leftarrow \text{clip}(\eta - \nabla_\eta \sum_{k=1}^{\mathcal{N}_\mathcal{P}} L_{A,k})$
  $\delta \leftarrow \delta - \nabla_\delta \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k})$
  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
  $L_B \leftarrow \sum_{j=1}^{\mathcal{N}_\mathcal{C}} L_{C,j} + \sum_{k=1}^{\mathcal{N}_\mathcal{P}} (\alpha L_{P,k} + \beta L_{M,k})$
  $\omega' \leftarrow \omega' - \nabla_{\omega'} L_B$
end for

datset, and derive the backdoor loss L_B by adding the clean loss, poison loss, and MSE loss with a fixed proportion. At last, the weight ω' of the model is updated by the gradient of the backdoor loss.

VI. Evaluation

A. Targeted Deep Learning Model

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 scheme. Existing work [3] has developed a Residual Neural Network (ResNet)-based classifier comprising five residual stacks to identify the modulation scheme of RF signals.

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. We implement the device identification model proposed by Sankhe *et al.* [4]. This model is designed based on Convolutional Neural Networks (CNNs), contains four layers including two convolution layers and two fully connected layers.

B. Experimental Methodology

RF Datasets. We employ the RF dataset collected by O'Shea *et al.* [3] 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 IQ 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.* [4] 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 backdoor 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 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 benign 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 Normalized Mean Squared Error (NMSE) [37] 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:

$$NMSE = \frac{MSE}{Var(x_c)} = \frac{\sum_{i=1}^{n}(x_{c_i} - x_{p_i})^2}{\sum_{j=1}^{n}(x_{c_j} - \bar{x}_c)^2}, \qquad (12)$$

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 A6000 GPUs. For the trigger initialization in an application, we calculate the variance σ_I^2 and σ_Q^2 corresponding to In-phase and Quadrature dimension in all RF IQ training sets. Then, we generate continuous IQ samples that follow the multivariate normal distribution $N(0, \sigma_I^2)$ and $N(0, \sigma_Q^2)$ as the initialized trigger. We establish the upper bound of the ratio of polluted IQ samples in a segment based on an empirical study by NMSE, which shows optimal outcomes can be achieved when the ratio of polluted samples is limited to 10%. NMSE improves as the ratio of polluted samples increases. Therefore, we conduct all our experiments with a max 10% polluting ratio. Besides, we poison max 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. The epoch size of 100 to avoid overfitting. The batch size is set to 1024 for both models. In the poison-label backdoor attack scenario, we empirically set the hyperparameters α and β to 0.3 and 0.2 in RF modulation classification. α and β are set to 0.2 and 0.05 in RF device identification. In the clean-label backdoor attack scenario, we empirically set the hyperparameters α and β to 0.2 and 0.1, 0.1 and 0.04 in two applications, respectively. The initialized adversarial perturbation is an unnoticeable Gaussian noise with the mean 0 and variance τ^2. The threhold of the adversarial perturbation τ is set to $0.001 \times \frac{(\sigma_I + \sigma_Q)}{2}$, which is also unnoticeable.

C. RF Backdoor Attack Performance

Poison-Label Attack Performance. The attack performance and clean data classification performance are presented in the TABLE I. We iteratively assign each of the labels as the

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3404341

TABLE I

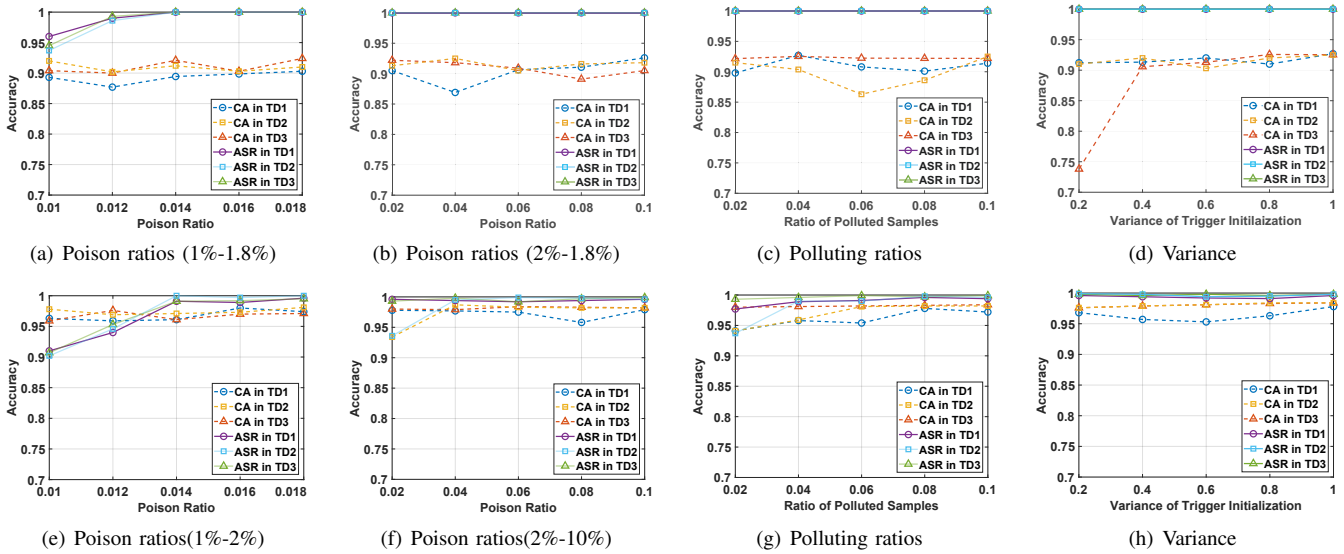
THE PERFORMANCE COMPARISON (ASR, NMSE, CA OF BACKDOORED MODEL AND CA OF BENIGN MODEL) OF THE RF POISON-LABEL BACKDOOR ATTACK WITH THE BACKDOOR TRIGGERS OPTIMIZED FOR RF MODULATION CLASSIFICATION AND RF FINGERPRINT-BASED DEVICE IDENTIFICATION.

Trigger Temporal Patterns	RF Modulation Classification			RF Fingerprint-based Device Identification		
	ASR	CA (Attack/Benign)	NMSE	ASR	CA ((Attack/Benign))	NMSE
TD1	100% (0.00%)	92.54%/92.50%	1.2×10^{-2}	99.28% (0.15%)	97.66%/98.19%	2.4×10^{-3}
TD2	100% (0.00%)	92.12%/92.50%	1.1×10^{-2}	99.55% (0.16%)	98.56%/98.19%	1.8×10^{-3}
TD3	100% (0.02%)	92.54%/92.50%	1.1×10^{-2}	99.61% (0.12%)	98.32%/98.19%	1.1×10^{-2}

TABLE II

THE PERFORMANCE COMPARISON (ASR, NMSE, CA OF BACKDOORED MODEL AND CA OF BENIGN MODEL) OF THE RF CLEAN-LABEL 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	CA (Attack/Benign)	NMSE	ASR	CA ((Attack/Benign))	NMSE
TD1	99.54%(0.11%)	93.48%/92.50%	9.1×10^{-2}	99.80%(0.10%)	98.08%/98.19%	6.9×10^{-2}
TD2	100%(0.00%)	92.61%/92.50%	8.4×10^{-2}	99.96%(0.03%)	98.32%/98.19%	6.6×10^{-2}
TD3	100%(0.01%)	93.40%/92.50%	8.7×10^{-2}	100%(0.01%)	97.74%/98.19%	6.2×10^{-2}



(a) Poison ratios (1%-1.8%) (b) Poison ratios (2%-1.8%) (c) Polluting ratios (d) Variance

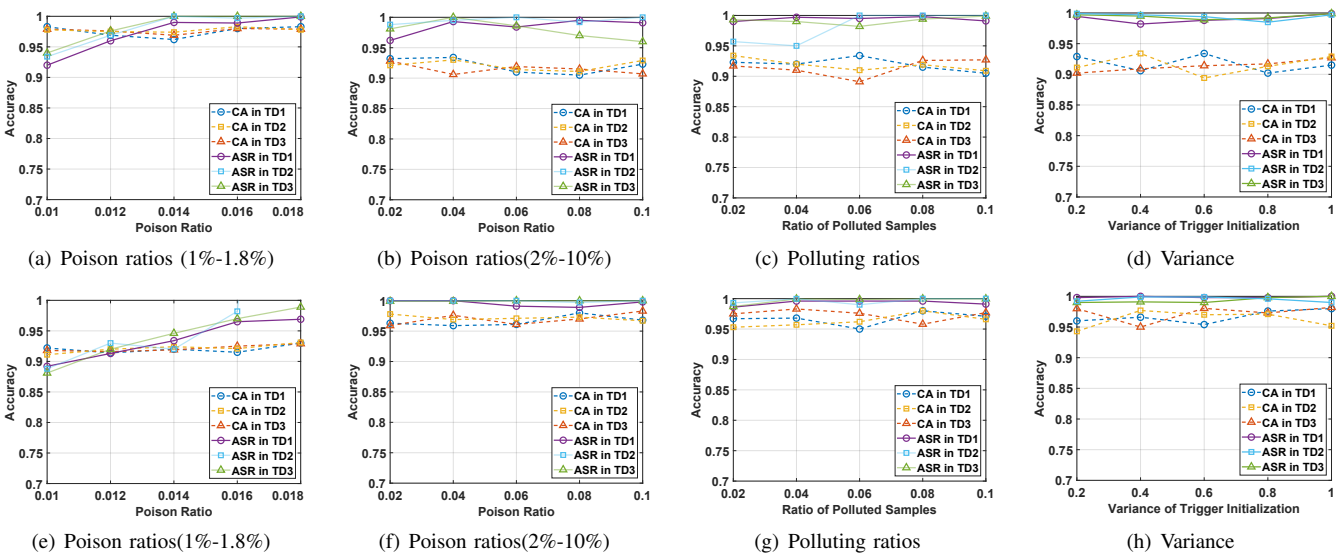
(e) Poison ratios(1%-2%) (f) Poison ratios(2%-10%) (g) Polluting ratios (h) Variance

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

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. The NMSE is less than 1.2×10^{-2} in modulation classification and 2.4×10^{-3} in device identification. These results demonstrate that our approach is stealthy since the poison samples polluted by our optimized trigger can be embedded in the clean samples. The performance in different patterns is similar as the prevalent RF signal classification model is CNN-based, which struggles to capture the temporal characteristics of data.

Clean-Label Attack Performance. The attack performance and clean data accuracy are presented in the TABLE II. We randomly assign three labels as the target labels and calculate the average ASR, CA, and standard deviation (STD) of ASR. In the application of RF modulation classification, the ASR can reach over 99.54% with low STD (less than 0.11%) across all trigger patterns. The clean data accuracy even increases when we launch the attack. In the radio frequency (RF) signal classification model, Gaussian noise is often incorporated into the raw data to enhance the model's accuracy. The trigger and adversarial perturbation are initialized by the Gaussian function, though optimized during the training process, also can be deemed as noise. The model that is trained with these noises is more robust to the subtle difference in one class. In the application of device identification, the ASR can reach over 99.80% with low STD (below 0.10%) while maintaining clean data accuracy (drop less than 0.45%). The NMSE of the trigger increases to over 0.084 in RF modulation classification and over 0.062 in RF device identification compared with the poison-label attack scenario. It's harder to implement

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3404341



(a) Poison ratios (1%-1.8%) (b) Poison ratios(2%-10%) (c) Polluting ratios (d) Variance

(e) Poison ratios(1%-1.8%) (f) Poison ratios(2%-10%) (g) Polluting ratios (h) Variance

Fig. 6. ASR and CA for RF modulation classification ((a), (b), (c),(d)) and RF fingerprint-based device identification (e), (f),(g),(h)) in clean-label backdoor attack with different impact factors (poison ratio, the ratio of polluted samples, and the variance of the Gaussian function in initial triggers).

a backdoor attack in a clean-label scenario as the label is unchanged, resulting in a higher NMSE. NMSE is below 0.2, which is far less than 1. The trigger is hard to be detected as an outlier. These results demonstrate that our optimized trigger can reach high attack performance and maintain clean data accuracy in the clean-label attack scenario.

D. Impacts Factors in Poison-Label Attack

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), (b), (e) and (f), 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., 1%, 1.2%, 1.4%, 1.6%, 1.8%, 2%, 4%, 6%, 8%, 10%). In RF modulation classification, the ASR can reach high and CA is stable by only poisoning 1.4% data. In RF device identification, The ASR can reach a high and stable value when the poisoning rate is over 1.4%. 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 (c) and Fig. 5 (f), we study the performance of our attack with different pollution ratios in a segment(i.e., 2%, 4%, 6%, 8%, 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 (d) and Fig. 5 (g), we investigate the performance of our backdoor attack with different ratios between initial variance and clean data 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 our backdoor attack algorithm is capable of learning from different initialized triggers to achieve similar optimal performance.

E. Impacts Factors in Clean-Label Attack

The performance is the same across multiple training runs as we fix the random seed, to study the impact of other impact factors. Using five different random seeds, the values of clean data accuracy, attack performance, and NMSE remain similar with the small std as 0.004, 0.001, and 0.009. respectively.
Impact of Poison Ratio. We study the performance with different poison ratios (i.e., 1%, 1.2%, 1.4%, 1.6%, 1.8%, 2%, 4%, 6%, 8%, 10%). As shown in Fig. 6 (a),(b),(e) and (f), our method can reach a high attack performance (over 99%) and maintain CA although in a low poison ratio with 1.4% in RF modulation classification. In RF device identification, ASR can reach high (over 99%) with a poison ratio 4%. Those results show the high efficiency of our method regarding the poison ratio of the training data for malicious tasks.

Impact of The Ratio of Polluted Samples. The length of the trigger is essential to the attack performance as a shorter trigger can be stealthier but challenging to reach high attack performance. We conduct our experiment with five ratios of polluted samples (i.e., 2%, 4%, 6%, 8%, 10%). As shown in Fig. 6 (c), regarding the modulation classification application, the ASR and CA remain stable as the increase of the ratio of polluted samples. In the application of device identification, Fig. 6 (g) has shown that the ASR remains stable but the CA floats slightly as the ratio of polluted samples increases.

Those results demonstrate that our method is efficient with a low ratio of polluted samples, which is stealthy.

Impact of The Variance in Trigger Initialization. We further evaluate the impact of the variance of the Gaussian function in trigger initialization by using five different ratios of variance and clean data variance (i.e., 0.2, 0.4, 0.6, 0.8, 1). As shown in Fig. 6 (d) and Fig. 6 (h), our method can reach optimal ASR and they remain stable with the increase of variance. Those results indicate that our method can reach optimal ASR and CA that are not significantly affected by the variance of the Gaussian function in trigger initialization.

F. Ablation Study

Impact of MSE Loss In Poison-Label Attack. In the poison-label attack scenario, we explore the effectiveness of MSE loss when training the model. We leverage MSE loss to constrain the amplitude of the trigger, making it imperceivable. As shown in Fig. 7 (a) and Fig. 7 (b), the NMSE of the model trained without MSE loss (over 0.8 in modulation classification and over 1.5 in device identification) is much higher than the model trained with MSE loss ($\leq$ 0.011 in modulation classification and $\leq$ 0.0024 in device identification) in three trigger patterns. The ASR can reach high (over 99%) with and without MSE loss. These results demonstrate that training with MSE loss can constrain the amplitude of the trigger while maintaining a high attack success rate.

Impact of MSE Loss In Clean-Label Attack. In the clean-label attack scenario, the NMSE and ASR of the model trained with and without MSE loss are illustrated in Fig. 7 (c) and Fig. 7 (d). The NMSE of the model trained without MSE loss is over 1.8 in modulation classification and over 2.4 in device identification of three trigger patterns. By involving MSE loss in the loss function, the NMSE drops to nearly 0.3 in modulation classification and less than 0.2 in device identification while ASR is maintained (around 96%), indicating the effectiveness of MSE loss to constrain the amplitude of trigger in the clean-label settings.

Impact of Adversarial Perturbation In Clean-Label Attack. To evaluate the performance of adversarial perturbation, we explore the NMSE and ASR of the model trained without and with adversarial perturbation. In the application of modulation classification shown in Fig. 7 (c), the NMSE is around 2.0 when training the model without adversarial perturbation and nearly 0.7 with adversarial perturbation. Meanwhile, the ASR increases to nearly 100% with adversarial perturbation. In the application of device identification shown in Fig. 7 (d), the NMSE is over 2.4 when training the model without adversarial perturbation and drops to around 0.5 with adversarial perturbation, while increasing ASR to over 99%. By training the backdoor model incorporated with MSE loss and adversarial perturbation, the NMSE further drops to less than 0.091 in modulation classification and less than 0.069 in device identification, while the ASR maintains high (nearly 100%). These results indicate the effectiveness of the adversarial perturbation to enhance the trigger

VII. Resistance to Backdoor Defense Methods

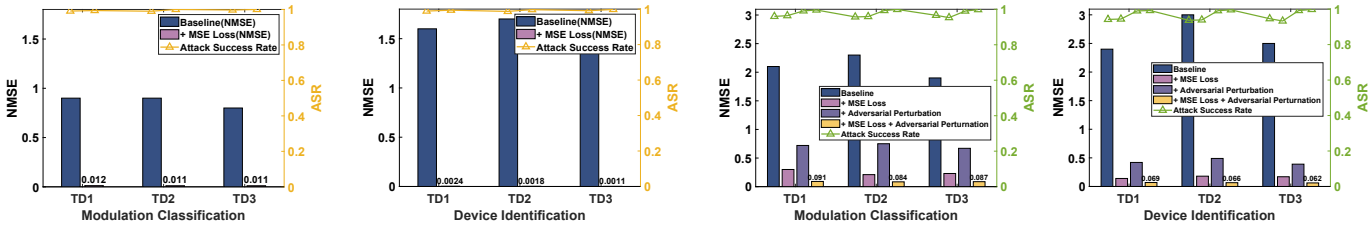
Resistance to Neural Cleanse. Neural Cleanse [38] is an optimization technique that reverse-engineers the model to detect the backdoor model. It works by searching for small perturbations that can cause the inputs to be classified into a particular class. The Anomaly Index is the primary metric used for detecting a backdoor model. When the value exceeds 2, it indicates the presence of a backdoor. As shown in Fig. 8 (a) regarding the poison-label attack scenario, we test our backdoor model in three temporal trigger patterns for two applications in the poison-label attack scenario. Anomaly index values for all those test cases are lower than 2. In the clean-label backdoor attack scenario, as shown in Fig. 8 (b), our method can also pass the detection. The results demonstrate that our attack is robust against Neural Cleanse.

Resistance to STRIP. STRIP [39] is an entropy-based backdoor detection approach to detect the presence of a backdoor by intentionally adding perturbations to the input and observing the probability distributions in the model's predictions. If the perturbed inputs are predicted as the same class, it leads to a low output distribution entropy, which signals a backdoor model's potential presence. We evaluate all three trigger patterns for two applications against STRIP. The results of the trained backdoor model of TD2 of RF modulation classification in two attack scenarios against STRIP are shown in Fig. 8 (c) and Fig. 8 (d). 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, indicating that the STRIP can't detect this backdoor model. We have similar observations for the trained backdoor models using other trigger designs on two applications. Those results show that our attack is robust against STRIP.

VIII. Conclusion

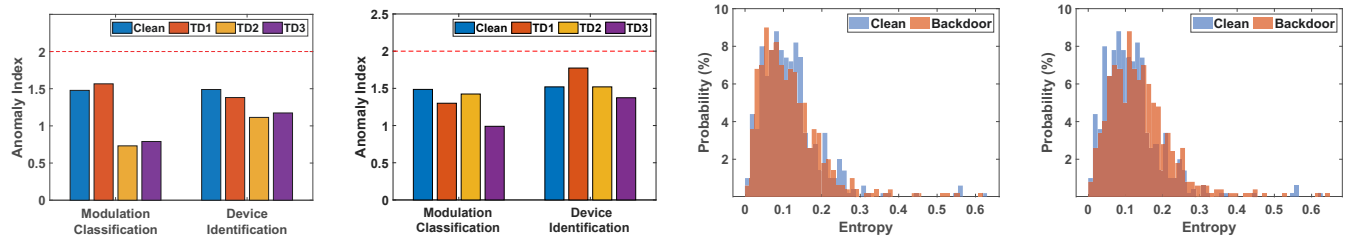
In this work, we propose the first stealthy RF backdoor attack that targets deep-learning-based RF signal classification applications. We thoroughly study the RF IQ data differences among different RF applications and within the same RF application. We find a fixed-positioned trigger can be easily detected. 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. We propose a training-based approach for generating triggers that can improve the efficiency and subtlety of RF signal classification. Furthermore, we develop two optimization techniques against to two backdoor attack scenarios: poison-label and clean-label, where the attacker may or may not have full control over the label when poisoning the data. 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.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3404341



(a) Modulation classification (b) Device identification (c) Modulation classification (d) Device identification

Fig. 7. NMSE and ASR for RF modulation classification and RF fingerprint-based device identification in poison-label attack ((a),(b)) without MSE loss (Baseline) and with MSE Loss (+MSE Loss) and NMSE and ASR in clean-label attack ((c),(d)) without MSE loss and adversarial perturbation (Baseline), only with MSE Loss (+MSE Loss), only with adversarial perturbation (+ Adversarial Perturbation), with MSE loss and adversarial perturbation (+MSE Loss and Adversarial Perturbation) in three trigger designs.



(a) Neural Cleanse (b) Neural Cleanse (c) STRIP (d) STRIP

Fig. 8. Illustration of the RF backdoor attack against two commonly used backdoor model detection approaches (i.e., Neural Cleanse and STRIP) in poison-label attack scenrio ((a),(c)) and clean-label attack scenario ((b),(d)).

ACKNOWLEDGMENT

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

REFERENCES

[1] T. Ulversoy, "Software defined radio: Challenges and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 4, pp. 531–550, 2010.

[2] T. J. OShea, 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.

[3] T. J. OShea, 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.

[4] 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.

[5] 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.

[6] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[7] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," *arXiv preprint arXiv:2007.10760*, 2020.

[8] H. C. Tanuwidjaja, R. Choi, S. Baek, and K. Kim, "Privacy-preserving deep learning on machine learning as a servicea comprehensive survey," *IEEE Access*, vol. 8, pp. 167 425–167 447, 2020.

[9] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2088–2105, 2020.

[10] X. Gong, Y. Chen, W. Yang, H. Huang, and Q. Wang, "b3: Backdoor attacks against black-box machine learning models," *ACM Transactions on Privacy and Security*, 2023.

[11] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 463–16 472.

[12] J. Ze, X. Li, Y. Cheng, X. Ji, and W. Xu, "Ultrabd: Backdoor attack against automatic speaker verification systems via adversarial ultrasound," in *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2023, pp. 193–200.

[13] 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.

[14] 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.

[15] 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.

[16] D. Roy, V. Chaudhury, C. Tassie, C. Spooner, and K. Chowdhury, "Icarus: Learning on iq and cycle frequencies for detecting anomalous rf underlay signals," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

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

[18] 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.

[19] A. Turner, D. Tsipras, and A. Madry, "Clean-label backdoor attacks," 2018.

[20] Y. Zeng, M. Pan, H. A. Just, L. Lyu, M. Qiu, and R. Jia, "Narcissus: A practical clean-label backdoor attack with limited information," *arXiv preprint arXiv:2204.05255*, 2022.

[21] N. Cauli, A. Ortis, and S. Battiato, "Fooling a face recognition system with a marker-free label-consistent backdoor attack," in *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*. Springer, 2022, pp. 176–185.

[22] N. Luo, Y. zhang Li, Y. Wang, S.-H. Wu, Y. Tan, and Q. xin Zhang, "Enhancing clean label backdoor attack with two-phase specific triggers," *ArXiv*, vol. abs/2206.04881, 2022.

[23] 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

[24] 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

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

[26] 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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 443–14 452.

[32] 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.

[33] 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.

[34] 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.

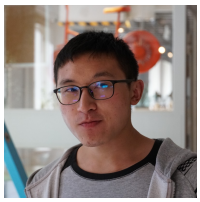
[35] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, and S.-T. Xia, "Backdoor attack against speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2560–2564.

[36] Y. Gao, Y. Li, L. Zhu, D. Wu, Y. Jiang, and S.-T. Xia, "Not all samples are born equal: Towards effective clean-label backdoor attacks," *Pattern Recognition*, vol. 139, p. 109512, 2023.

[37] A. A. Poli and M. C. Cirillo, "On the use of the normalized mean square error in evaluating dispersion model performance," *Atmospheric Environment. Part A. General Topics*, vol. 27, no. 15, pp. 2427–2434, 1993.

[38] 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.

[39] 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.



Zijie Tang received the B.E degree in Computer Science and Technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2022. He is currently working toward the Ph.D. degree at the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. He is currently working with Prof. Yan Wang. His research interests include machine learning and smart sensing.

Tianming Zhao received Ph.D. degree at the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA, in 2022, the M.S. degree from the Department of Computer Science, Binghamton University, Binghamton, NY, USA, in 2016 and the B.E. degree from the Department of Software Engineering, Chongqing University, Chongqing, China, in 2013. He is currently an Assistant Professor with the Department of Computer Science, University of Dayton, Dayton, Ohio, USA. His research interests include mobile computing and sensing, cybersecurity and privacy, and smart health.

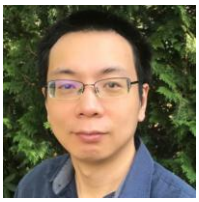
Tianfang Zhang received the B.E degree in information security from the Huazhong University of Science and Technology, Wuhan, China, in 2019, and the M.S. degree in computer science in 2021 from Rutgers University, New Brunswick, NJ, USA. He is currently working toward the Ph.D. degree in electrical and computer engineering at Rutgers University, New Brunswick, NJ, USA. He is currently working with Prof. Yingying Chen. His research interests include machine learning, mobile system, wireless sensing, and cyber security.

Huy Phan received the B.E degree in Computer Science from Rutgers University, New Brunswick, NJ, USA, in 2018, and the M.S. degree in Electrical and Computer Engineering in 2020 from Rutgers University, New Brunswick, NJ, USA. He is currently working toward the Ph.D. degree in electrical and computer engineering at Rutgers University, New Brunswick, NJ, USA. He is currently working with Prof. Bo Yuan. His research interests include computer vision, deep learning and deep learning sercurity.

Yan Wang received the Ph.D. degree in electrical engineering from the Stevens Institute of Technology, Hoboken, NJ, USA, in 2015. He was with the Department of Computer Science, State University of New York (SUNY), Binghamton, NY, USA. He is currently an Assistant Professor with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. His research interests include cybersecurity and privacy, the Internet of Things, mobile and pervasive computing, and smart healthcare. His research is supported by the National Science Foundation (NSF). His research has been reported in numerous media outlets, including MIT Technology Review, CNN, Fox News Channel, The Wall Street Journal, National Public Radio, and IEEE Spectrum. Dr. Wang was a recipient of the Best Paper Award from IEEE CNS 2018, IEEE SECON 2017, and ACM AsiaCCS 2016. He was the winner of the ACM MobiCom Student Research Competition in 2013.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3404341

Cong Shi recieved the Ph.D. degree with Wireless Information Network Lab, Rutgers University, New Brunswick, NJ, USA. He was the recipient of the two industry-sponsored fellowships by Cisco System and Siemens Corporate Research. He is currently an Assistant Professor with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. His research interests include cyber security and privacy, mobile sensing, smart healthcare, Internet of Things, security in machine learning, and artificial intelligence.



Bo Yuan received the B.E and M.E degrees from Nanjing University, Nanjing, China, in 2007 and 2010, respectively, and the PhD degree from the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities, Minnesota, in 2015. His research interests include algorithm and hardware co-design and implementation for machine learning and signal processing systems, error-resilient low-cost computing techniques for embedded and IoT systems, and machine learning for domain-specific applications. He is the recipient of Global Research Competition Finalist Award in Broadcom Corporation and doctoral dissertation fellowship with the University of Minnesota. He serves as technical committee track chair and technical committee member for several IEEE/ACM conferences. He is the associated editor of the Springer Journal of Signal Processing System.

Yingying Chen (Fellow, IEEE) is currently a professor of electrical and computer engineering and Peter Cherasia endowed faculty scholar with Rutgers University. She is the associate director of Wireless Information Network Laboratory (WINLAB). She also leads the Data Analysis and Information Security (DAISY) Lab. She is a NAI fellow. She is also named as an ACM distinguished scientist. Her research interests include mobile sensing and computing, cyber security and privacy, Internet of Things, and smart healthcare. She is a pioneer in the area of RF/WiFi sensing, location systems, and mobile security. She had extensive industry experience with Nokia previously. She has published three books, four book chapters and more than 200 journal articles and refereed conference papers. She is the recipient of seven best paper awards in top ACM and IEEE conferences. Her research has been reported by numerous media outlets. She has been serving/served on the editorial boards of the IEEE/ACM Transactions on Networking (IEEE/ACM ToN), IEEE Transactions on Mobile Computing (IEEE TMC), IEEE Transactions on Wireless Communications (IEEE TWireless), and ACM Transactions on Privacy and Security.