W-MIA: Membership Inference Attack against Deep Learning-based RF Fingerprinting

Yan Zhang*, Jiawei Li*, Dianqi Han†, Yanchao Zhang*
* Arizona State University, † University of Texas at Arlington
{zhangyan, jwli, yczhang}@asu.edu, dianqi.han@uta.edu

2024 IEEE Conference on Communications and Network Security (CNS) | 979-8-3503-7596-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/CNS62487.2024.10735522

***Abstract*—Deep learning-based RF fingerprinting (DRFF) systems have gained prominence for their effectiveness in wireless device authentication based on unique RF hardware features in wireless signals. However, the inherent vulnerabilities of deep learning (DL) models make DRFF systems susceptible to DL attacks tailored for RF fingerprinting. In this paper, we present W-MIA, the first practical label-only membership inference attack (MIA) against DRFF systems. W-MIA can passively eavesdrop on RF signals to construct a shadow model and perform MIA covertly. Additionally, it can enhance attack efficacy through low-rate tailored active interactions with DRFF systems. We also propose a simple yet effective countermeasure against W-MIA. Extensive experiments confirm W-MIA's high attack efficacy in a label-only setting, achieving a maximum AUC of 0.81, comparable to the latest MIA against DRFF, which assumes a more knowledgeable adversary. Furthermore, our proposed defense matches the performance of existing defenses while minimizing usability loss in DRFF systems.**

***Index Terms*—RF fingerprinting, deep learning, membership inference attack, wireless security.**

I. Introduction

Deep Learning-based RF Fingerprinting (DRFF) [1]–[4] is promising for wireless device authentication by utilizing unique signal distortions caused by on-device RF chip imperfections. Fig. 1 shows a typical DRFF system that comprises two phases. In the enrollment phase, a system server collects RF Fingerprinting (RFF) samples from each legitimate device to form a training dataset. These samples are then used to train a DRFF-DNN model, which is a multi-class classifier with each class corresponding to a unique device. In the authentication phase, a *verifier* employs the DRFF-DNN model to classify the RFF samples emitted by a *claimant*, determining whether they match the alleged device. Only upon successful authentication can the claimant gain access to protected system resources or engage in subsequent communications with the verifier. The verifier can take various forms in different application contexts, such as an access point, gate-control device, or even a regular device similar to the claimant. Since RFF samples are extracted from standard wireless transmissions, the DRFF-based authentication process can be one-time or continuous, and one-way or bidirectional.

In this paper, we investigate the vulnerability of DRFF to the Membership Inference Attack (MIA) [5]–[8], a known privacy threat against deep learning (DL) models. In the DRFF context, an adversary launches the MIA to determine if an RFF sample, captured from a legitimate device during the authentication phase, is present in the training dataset of the target DRFF-DNN model. During authentication, legitimate devices emit RFF samples that include both training and non-training samples, hereafter referred to as "member" and "non-member" samples of the training dataset, respectively. These samples can vary due to factors such as channel conditions, device locations, and measurement time. However, all can be classified into the same legitimate device by the well-trained DRFF-DNN model. The MIA is feasible because the DRFF-DNN model tends to overfit member RFF samples, resulting in observable classification outcomes visible to the adversarial sniffer. A successful MIA poses two primary threats to DRFF-based wireless authentication: 1) enabling attackers to mimic member RFF samples to bypass the authentication system [3], [9], and 2) significantly improving the success rate of other attacks [10]–[13]. For example, since member samples are farther away from the decision boundary than non-member samples in the DRFF-DNN model, the adversary has a higher chance of successfully impersonating a legitimate device by mimicking its member RFF sample inferred under the MIA. This is because the adversary's well-crafted mimicry of a member RFF sample is more likely to remain within the

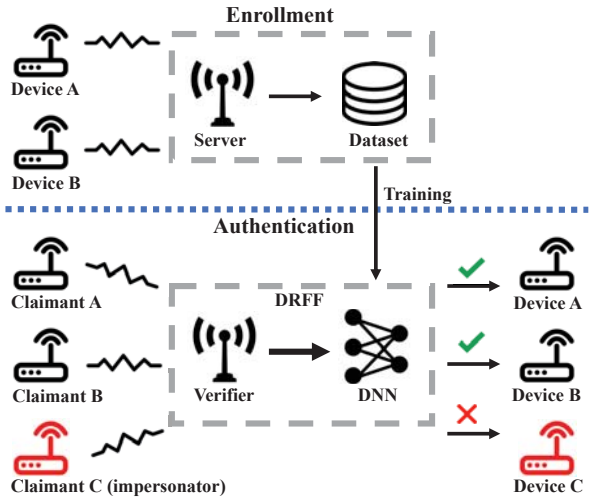


Fig. 1: DRFF-based wireless authentication.

victim device's decision region in the DRFF-DNN model, even when influenced by dynamic wireless channels.

Implementing a practical MIA against DRFF is notably challenging due to three factors. Firstly, attackers often lack comprehensive knowledge of the DRFF-DNN model's architecture and parameters. Secondly, to avoid detection, attackers must limit their forged inquiries to the DRFF system. Thirdly, such requests typically yield basic "label-only" information, indicating merely the (il)legitimacy of the claimant. The attack reported in [14] is the only MIA targeting DRFF. However, this attack assumes that the adversary knows the classification scores output by the DRFF-DNN model for each claimant under sniffing, rendering it ineffective when this assumption does not hold.

To address these challenges, we propose W-MIA, a practical MIA against DRFF systems. W-MIA creates an imitation of the target DRFF-DNN model, referred to as a *shadow model*, which is used with classic MIA techniques to infer membership information. The attacker obtains the shadow model through two steps: *Passive* W-MIA (Step 1) and *Active* W-MIA (Step 2). Initially, the attacker passively observes the verifier's responses to authentication requests and trains a preliminary shadow model accordingly. Then, the attacker actively forges re- quests to the verifier to gather additional information for refining the shadow model. Active W-MIA involves adver- sarial perturbation generation for request forgery, allowing the attacker to obtain comprehensive information on the DRFF-DNN model from each authentication request. This approach results in a high-quality shadow model with minimal requests, making the MIA more stealthy. In contrast to the state of the art [14], W-MIA only assumes that the adversary can passively observe the (il)legitimacy of a claimant given by the DRFF-DNN model, based on subsequent communications between the verifier and the claimant. This assumption is relatively easy to satisfy, making W-MIA a more practical threat against DRFF.

We evaluate W-MIA using a simulated DRFF system and a public RF fingerprint dataset. W-MIA achieves AUC scores of 0.75 and 0.81 in passive and active modes, respectively. These scores are comparable to the AUC score of 0.83 achieved by the prior work [14], which nevertheless assumes a much more informed attacker.

We propose *selective response* as a simple yet highly effective countermeasure against W-MIA. It functions by allowing the verifier to selectively ignore requests from enrolled devices. This disrupts the attacker's information gathering in both steps of W-MIA, thereby preventing the training of a high-quality shadow model. Additionally, we analyze the outputs of the DRFF-DNN model to identify requests with minimal impact on system usability. Our evaluation confirms the effectiveness of selective response. It successfully reduces the AUC of active W-MIA from 0.81 to 0.69, accompanied by only a slight increase in the false-negative rate (FNR) from 2% to 9%. In contrast, some known MIA defenses, such as adversarial regularization (AR) [15], increase the FNR to 21% or more to mitigate W-MIA.

The paper is structured as follows. §II presents sys- tem and adversary models. §III and §IV detail passive and active W-MIA designs, respectively. §V presents our countermeasures against W-MIA. §VI demonstrates experimental results. §VII discusses related studies. §VIII concludes this work.

II. System and Adversary Models

A. System Model

W-MIA targets a general DRFF-based wireless authentication system, as depicted in Fig. 1. This system consists of an enrollment phase, where the RF fingerprints of legitimate devices are registered, followed by an authentication phase, where a device's RF fingerprint is verified to confirm its alleged identity. While W-MIA can work with arbitrary RF fingerprints, for ease of illustration, we assume WiFi-based RF fingerprints in this paper.

Enrollment Phase. This phase occurs in a controlled and secure environment without potential attackers. Let's assume that N legitimate devices are enrolled in the system. Each device, denoted by $i \in [1, N]$, transmits a specified number of regular WiFi packets from different locations and at various times to the system server. In existing DRFF systems [1], [3], [9], [16], the server commonly extracts the complex-valued I/Q data containing the preamble of each captured WiFi packet and tags them with the device ID (MAC address). Either the raw I/Q data or their enhancements to minimize channel-dependent features [1], [3], [9] carry the unique RF fingerprint of the corresponding device and are referred to as an RFF sample for DRFF systems. Let D_i denote the set of RFF samples collected from device i during enrollment. The training dataset is then denoted by $\mathcal{D} = \bigcup_{i=1}^{N} D_i$, where each RFF sample is called a member sample.

The server then uses the training dataset $\mathcal{D}$ to train a DRFF-DNN model, denoted as $\mathcal{M}$, employing any feasible DNN architecture such as those used in [1], [3], [9], [16]. $\mathcal{M}$ operates as a multi-class classification system, which provides either an enrolled device ID or "unknown" as output for any input RFF sample.

Authentication Phase. Each authentication session involves a verifier and a claimant, where the claimant's alleged device ID needs verification. The claimant can be any WiFi device, while the verifier may be an access point, gate-control device, or a WiFi device similar to the claimant. It is assumed that the verifier has securely acquired the DRFF-DNN model $\mathcal{M}$ from the server. From each WiFi packet sent by the claimant, the verifier extracts one RFF sample, which is then processed by $\mathcal{M}$. If $\mathcal{M}$ outputs a device ID that matches the alleged one, the claimant passes the authentication; otherwise, it is considered illegitimate. Only legitimate claimants are allowed

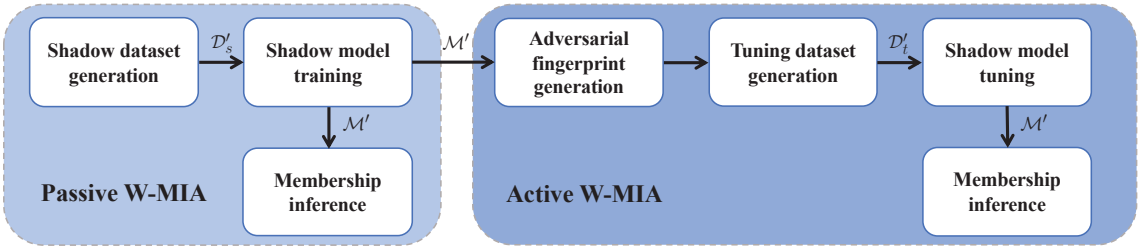


Fig. 2: Two-stage W-MIA diagram.

to enter a gated facility, access network resources, or engage in subsequent communications with the verifier. An authentication session may fail due to signal distortions or model classification errors, in which case the claimant typically resubmits its authentication request up to a certain threshold. Since RFF samples can be verified on a per-packet basis, the authentication session can be one-time or continuous. Authentication can also be bidirectional, with the verifier and claimant switching roles. *For clarity, we assume a very useful scenario where the verifier acts as an access point, continuously authenticating RFF samples from the claimant throughout the communication session.*

B. Adversary Model

W-MIA aims to determine whether an RFF sample belongs to the training dataset $\mathcal{D}$. We make the following assumptions about the attacker's capabilities. First, the attacker can place a spy RF monitor near the verifier to intercept WiFi packets during authentication. This is feasible because, unlike the enrollment phase which occurs in a controlled and secure environment, the authentication phase typically takes place in an open and insecure wireless environment, providing ample opportunities for adversarial packet interception and injection. Second, the attacker can decode and access unencrypted data, especially the device ID, exchanged between the verifier and the claimant. Third, the attacker can generate arbitrary WiFi packets with minimal distortion using software-defined radios [3]. Additionally, the attacker has label-only access to $\mathcal{M}$: they can impersonate legitimate devices to submit forged authentication requests encompassing various RFF samples to the verifier and observe their responses. However, the attacker lacks knowledge of $\mathcal{M}$'s architecture and internal parameters, thus implementing a label-only black-box MIA [5]–[8].

III. Passive W-MIA

In this section, we present the design of passive W-MIA, the initial phase of W-MIA. Passive W-MIA involves compiling a shadow dataset $\mathcal{D}'_s$ to train a shadow model $\mathcal{M}'$, which mimics the target DRFF-DNN model $\mathcal{M}$. Subsequently, the attacker may optionally perform membership inference to gain knowledge about $\mathcal{M}$.

A. Shadow Dataset Generation

We employ a passive, query-free method to generate $\mathcal{D}'_s$. Our method capitalizes on the fact that the verifier typically responds only when $\mathcal{M}$ successfully validates the association of the embedded RFF sample in the claimant's message with the alleged device ID. Consider an arbitrary claimant, say the enrolled device i, whose ID is denoted by l_i. Assume that the attacker can deploy a passive sniffer near the verifier to intercept all the packets between the verifier and claimant i. Since the packet preamble and l_i are usually unencrypted, the attacker can easily extract one RSS sample, denoted by f_i, from each packet arriving from device i. If the verifier responds to claimant i's packets, the attacker adds $\langle f_i, l_i \rangle$ to $\mathcal{D}'_s$. Otherwise, the attacker discerns that the DRFF-DNN model $\mathcal{M}$ classifies f_i as "unknown" or associates it with an enrolled device other than l_i. Without further information, the attacker simply adds $\langle f_i, \text{"unknown"} \rangle$ to $\mathcal{D}'_s$.

There are two important points to note. First, the attacker only needs to sniff RSS samples from the legitimate devices they are interested in targeting with the MIA. In other words, $\mathcal{D}'_s$ does not need to include RSS samples from all legitimate devices. Second, $\mathcal{D}'_s$ may contain both member and non-member RSS samples of the targeted legitimate devices. Our evaluations in §VI show that the attacker only needs to sniff a small percentage of member RSS samples from a legitimate device, or even only its non-member samples, for a sufficiently effective MIA.

B. Shadow Model Training

The attacker proceeds to build the shadow model $\mathcal{M}'$ using $\mathcal{D}'_s$. Despite being unaware of the exact architecture of the target model $\mathcal{M}$, the attacker can still utilize DNN architectures proven effective in RF fingerprinting. In this study, we assume that the attacker uses $\mathcal{D}'_s$ to train three candidate shadow models based on Homegrown [1], Oracle [3], and ResNet [1], respectively. The candidate model with the highest accuracy is chosen as $\mathcal{M}'$.

C. Membership Inference on $\mathcal{M}'$

The attacker can perform membership inference with $\mathcal{M}'$ to gain insights into $\mathcal{M}$. This process is akin to a white-box W-MIA because the attacker has complete knowledge of $\mathcal{M}'$. In W-MIA, the attacker uses the cross-entropy loss [6] to distinguish between member and non-member RSS samples for $\mathcal{M}$ with $\mathcal{M}'$ instead. Typically, $\mathcal{M}$ exhibits higher confidence in classifying member RSS samples, resulting in lower cross-entropy loss values. Conversely, the loss is higher for non-member RSS samples. We expect

the same result with its close imitation, $\mathcal{M}'$. To determine if a newly sniffed RSS sample $\langle f_t, l_t \rangle$ is a member sample of $\mathcal{M}$, the attacker inputs f_t into $\mathcal{M}'$ and calculates its cross-entropy loss. $\langle f_t, l_t \rangle$ is classified as a member sample if its cross-entropy loss value is below a predetermined threshold, and as a non-member sample otherwise.

IV. Active W-MIA

Active W-MIA involves refining the shadow model $\mathcal{M}'$ to reduce its differences from $\mathcal{M}$, thereby enhancing the accuracy of membership inference. In this process, the attacker actively submits fabricated packets containing specially crafted RFF samples to the verifier and adjusts $\mathcal{M}'$ based on the verifier's responses. The refined shadow model is still referred to as $\mathcal{M}'$ for simplicity. Active W-MIA involves three main steps illustrated below.

A. Adversarial RFF Sample Generation

Effective model tuning requires RFF samples that exhibit different classifications between $\mathcal{M}'$ and $\mathcal{M}$ and are near the decision boundaries of both models. We first obtain some candidate samples through the Basic Iterative Method (BIM) [17], a popular method for adversarial perturbation generation that demonstrates superior performance in our evaluations compared to other methods.

For each RFF sample $f \in \mathcal{D}'_s$, the attacker aims to generate a minimal perturbation ϵ such that $\mathcal{M}'(f+\epsilon) = l_a$, where $l = \mathcal{M}'(f)$, and l_a represents every legitimate device class other than l. We refer to $f+\epsilon$ as an adversarial RFF sample, denoted as f_a. Specifically, the attacker iteratively adjusts ϵ until $f+\epsilon$ is classified as the target class l_a by $\mathcal{M}'$. In each iteration, the attacker updates ϵ as follows:

$$\epsilon := \epsilon - \alpha \cdot \text{sign}(\nabla_f \mathcal{L}(f+\epsilon, l_a)). \tag{1}$$

Here, $\mathcal{L}(\cdot)$ represents the loss function used in training $\mathcal{M}'$. The term $\nabla_f \mathcal{L}(f+\epsilon, l_a)$ computes the gradient of the loss function with respect to the input to $\mathcal{M}'$. The sign$(\cdot)$ function creates a matrix with elements set to 1 or -1, corresponding to the signs of the elements in the input matrix. The scaling factor α is fixed at 0.0001, an empirical value found to be very effective in our evaluations. Importantly, if the attacker cannot achieve an effective perturbation within a threshold number of iterations (40 in our evaluations), they consider the attempt unsuccessful and proceed to the next instance, where another legitimate device class other than l becomes the target class for l_a. Given N legitimate devices in the system, the adversary can generate up to $N-1$ adversarial RFF samples for each RFF sample in the shadow dataset $\mathcal{D}'_s$.

B. Tuning Dataset Generation

We then employ a two-step process to generate a tuning dataset, denoted by $\mathcal{D}'_t$, by selecting adversarial samples classified differently by $\mathcal{M}'$ and $\mathcal{M}$. First, the attacker estimates the channel condition, denoted as H, to the verifier by collaborating with their spy monitor near the verifier. This step is crucial to mitigate the channel impact on the RFF sample extracted by the verifier. Second, for each adversarial sample $\langle f_a, l_a \rangle$, the attacker impersonates device l to submit an authentication request, embedding f_a/H as the RFF sample, where l denotes the original device ID (as classified by $\mathcal{M}'$) used to generate $\langle f_a, l_a \rangle$. If the verifier responds as if receiving a legitimate request, the attacker can infer that $\mathcal{M}$ associates f_a with the alleged device l, thus adding $\langle f_a, l \rangle$ to $\mathcal{D}'_t$. Otherwise, $\mathcal{M}$ either associates f_a with devices other than l or classifies it as "unknown", in which case f_a is excluded from $\mathcal{D}'_t$.

C. Shadow-Model Tuning

Finally, the attacker fine-tunes the shadow model $\mathcal{M}'$ using $\mathcal{D}'_t$ to further reduce its dissimilarity from $\mathcal{M}$. The model-tuning process consists of multiple epochs. Within each epoch, the attacker iteratively employs every data record in $\mathcal{D}'_t$ to adjust the parameters of $\mathcal{M}'$. Specifically, for a data record $\langle f_t, l_t \rangle$, the internal parameters of $\mathcal{M}'$ are updated as follows:

$$\theta'_t = \theta'_t + \lambda \cdot \nabla_{\theta'_t} \mathcal{L}(\theta'_t, f_a, l_i). \tag{2}$$

Here, θ'_t represents the internal parameters of $\mathcal{M}'$. $\nabla_{\theta'_t} \mathcal{L}(\theta'_t, f_a, l_i)$ computes the gradients of θ'_t with respect to the loss value. The learning rate λ and the total number of epochs are empirically determined through experiments, as detailed in §VI.

V. Countermeasure

We propose *selective response* to defend against W-MIA. The core idea is simple yet effective: the verifier deliberately responds oppositely to certain RFF samples, introducing confusion for potential attackers. Specifically, the verifier ignores some valid RFF samples to induce noise in the attacker's shadow and tuning datasets, thus hindering the development of an accurate shadow model.

To minimize disruption to the normal wireless authentication process, our system only rejects RFF samples that incur high-loss values upon validation by the target DRFF-DNN model $\mathcal{M}$. Prior research indicates that RF fingerprints validated by $\mathcal{M}$ with low loss values typically occur in stable, low-noise settings [3]. Rejecting such requests could lead to the immediate resubmission of a very similar RFF sample from the same authorized device. Since the subsequent RFF sample provides comparable data for membership inference, it too should be rejected. If this process continues, it might result in the unjustified denial of access to the enrolled device for a prolonged duration. To mitigate this, we propose discarding only those valid requests whose RFF samples achieve a loss value exceeding a predefined threshold η. In deep learning, these high-loss RFF samples are often indicative of outliers, mislabeled data, or particularly challenging cases for the model to learn. It is thus more difficult for the adversary to distinguish those deliberately rejected legitimate high-loss RFF samples from truly illegitimate ones. The determination of η is based on an experiment demonstrated in §VI.

VI. Performance Evaluation

A. Experimental Setup

1) Datasets: We use two datasets to evaluate W-MIA: a MATLAB-simulated dataset denoted by D_1 and a public dataset denoted by D_2 [1].

MATLAB-simulated Dataset D_1. We simulate a DRFF-based wireless authentication system using the MATLAB Communication Toolbox. Our simulation involves 21 virtual RF devices: one designated as the verifier, 19 as legitimate devices, and an additional one that mimics an illegitimate device generating a substantial number of negative RFF samples. Each virtual device, except the verifier, is configured with a unique RF fingerprint, represented by its phase shift, carrier frequency offset, and DC offset. These parameters for each virtual device are randomly sampled from [0.01°, 0.3°] for phase shift, [-4 ppm, 4 ppm] for carrier frequency offset, and [-50 dBc, -32 dBc] for DC offset. We use the Rayleigh channel model to simulate the multipath channel between each device and the verifier. Our simulations involve one Line-of-Sight (LOS) path and two Non-Line-of-Sight (NLOS) paths. Each path is assigned specific propagation delays and path gains. To create a dynamic environment, we randomly select the path delays and gains for each transmission between each device and the verifier.

Using this MATLAB simulator, we collect 5,000 RSS samples from each non-verifier virtual device and 100,000 in total, corresponding to their member samples. These 100,000 samples constitute the dataset used to train the target DRFF-DNN model $\mathcal{M}$. Furthermore, we employ the same simulator for the authentication phase of the wireless authentication system for passive W-MIA evaluation. For the evaluation of active W-MIA, we also use this simulator to simulate the request-forgery process.

Public Dataset D_2. We also evaluate W-MIA with a public dataset [1], collected in a real RFF system based on USRP devices. From this dataset, we select 21 devices: one serving as the verifier, 19 as legitimate devices, and the last one as an unknown device that generates negative RFF samples. During data collection, each non-verifier device transmits the same WiFi frame containing 288 I/Q samples, which form one RFF sample of the device when received by the verifier. The dataset for 20 non-verifier devices comprises 8.05×10^6 RFF samples collected in various environmental settings. We randomly select 100,000 samples for training the target model $\mathcal{M}$, including 5,000 from each device. It's important to note that this dataset is suitable only for passive W-MIA evaluation.

2) Target DRFF-DNN models: Our evaluation uses three DRFF-DNN model architectures that have been proven effective: Homegrown [1] (C_1), modified Oracle [3] (C_2), and ResNet [1] (C_3). We use datasets D_1 and D_2 to train these models, generating six target models in total. The model training is performed using TensorFlow and involves cross-validation to prevent overfitting.

3) Performance metric: The main performance metric we use is the AUC score, commonly employed in membership inference assessment. It measures the area under the ROC curve, depicting the relationship between the true positive rate (TPR) and the false positive rate (FPR). The AUC score ranges from 0.5 to 1, with higher scores indicating better performance.

We follow the following steps to calculate the AUC scores. For the evaluation of W-MIA on a target model, we generate an evaluation dataset comprising 1,000 mem- ber RFF samples of the model and 500 non-member RFF samples. W-MIA utilizes cross-entropy for mem- bership inference. Consequently, we examine various en- tropy threshold values and compute the corresponding true positive rate (TPR) and false positive rate (FPR), defined as $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ and $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$, respec- tively. Here TP represents the number of member samples correctly identified; FP represents the number of non-member samples incorrectly identified as members; TN represents the number of non-member samples correctly identified as non-members; and FN represents the number of member samples incorrectly identified as non-members. Subsequently, we use the scikit-learn Python package to derive the AUC score.

B. General Experimental Process

This section outlines a general process for implementing and evaluating W-MIA. We apply W-MIA to different settings, encompassing both passive and active W-MIA.

1) Passive W-MIA evaluation: Given a target model $\mathcal{M}$, we collect the shadow dataset as follows. For $\mathcal{M}$ trained on dataset D_1, we utilize the MATLAB simulator to simulate authentication phases using 20 virtual devices, generating RFF samples under varying channel conditions. For $\mathcal{M}$ trained on dataset D_2, we randomly select RFF samples for each device from the original dataset [1], which are not part of D_2. In both cases, only RFF samples that pass $\mathcal{M}$'s classification are considered non-member samples of the corresponding device and are consequently added to the shadow dataset. Since the shadow dataset does not include any member samples, it amounts to evaluating the worst-case performance of W-MIA.

We then train two candidate shadow models using different DNN architectures from the target model $\mathcal{M}$. For example, if $\mathcal{M}$ uses architecture C_1, we create candidate shadow models based on C_2 and C_3. This aligns with the assumption that a label-only attacker is unlikely to know $\mathcal{M}$'s exact architecture. The candidate model with the best accuracy on the shadow dataset is selected as the final shadow model. Finally, we perform a white-box MIA on the shadow model and calculate its AUC score.

2) Active W-MIA evaluation: We evaluate active W-MIA using a MATLAB simulator, excluding the public dataset because it cannot support request forgery. We initiate the process by generating adversarial perturbations

on RFF samples within the shadow dataset, resulting in a collection of adversarial RFF samples.

Next, we simulate the RF channel between the attacker and verifier to enable request forgery. For each adversarial RFF sample, we simulate a channel using six randomly selected parameters, including delays and gains for the three communication paths, to model the interaction between the attacker and verifier. The resulting channel state serves as the outcome of the channel sounding process. To introduce channel variations, we perturb the six channel parameters, obtaining the actual channel state when the adversarial RFF sample is transmitted. Each parameter is scaled by a factor of $1 + r$, where r is a random value within the range [-b, b]. We explore different values of b to simulate various levels of environmental variation. The collected adversarial RFF samples, along with their corresponding device classes classified by $\mathcal{M}$, compose the tuning dataset. Finally, we fine-tune the shadow model using this dataset and conduct the white-box MIA to calculate its AUC score.

C. Evaluation Results for Passive W-MIA

This section evaluates the impact of shadow datasets on passive W-MIA and experimentally demonstrates the significance of shadow model selection.

1) Impact of shadow dataset size: The size of the shadow dataset is critical for W-MIA performance, so we assess its impact through experiments. Specifically, we conduct passive W-MIA on six target models, corresponding to the two datasets and three architectures. For each target model, we collect shadow datasets of varying sizes to perform the attack, and the corresponding AUC scores are shown in Fig. 3. Generally, a larger shadow dataset results in a higher AUC score. Notably, a shadow dataset comprising 20,000 RSS samples can achieve an AUC score higher than 0.7, a commonly used threshold to indicate effective membership inference [6]. Moreover, further increasing the shadow dataset size beyond a certain point (50,000 for D_1 and 35,000 for D_2) does not yield significant improvement. Therefore, we have chosen to use a shadow dataset size of 35,000 for all the subsequent experiments in this section.

We also highlight the practicality of sniffing enough RSS samples for passive W-MIA. Take 802.11n as an example. A normal WiFi transmission speed typically ranges from about 6,250 to 25,000 frames per second. In our target scenario, where the verifier continuously verifies the RFF of an alleged device on a per-frame basis or every few frames, it would take the adversary well below one second to overhear enough RFF samples from any target victim device to establish the shadow dataset.

2) Impact of shadow model selection: This experiment aims to confirm the significance of shadow model selection. We conduct passive W-MIA on six target models and evaluate the AUC scores and classification accuracy with their respective shadow dataset. In addition, we evaluate MLP and Random Forest as candidate shadow models for comparison, denoted as C_4 and C_5, respectively. The results are presented in Fig. 4. As expected, shadow models designed specifically for RFF significantly outperform those based on general ML architectures. Furthermore, there is a strong correlation between the candidate shadow model's accuracy and its AUC score. Specifically, candidate models with the highest accuracy also achieve the best AUC scores, supporting our selection principle.

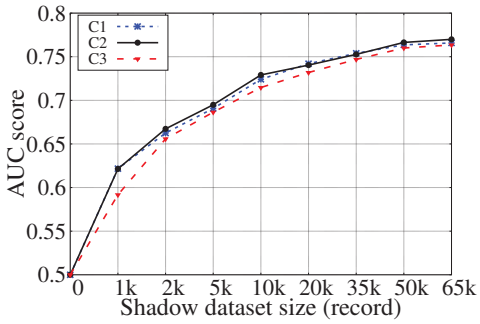
D. Evaluation Results for Active W-MIA

1) BIM vs. FGSM for adversarial perturbation: The most popular methods for adversarial perturbation gen- eration include the Fast Gradient Sign Method (FGSM) [18] and BIM [17]. We conduct an experiment to justify our choice of BIM over FGSM for generating adversarial RFF samples in active W-MIA. Specifically, we select the model trained on dataset D_1 with architecture C_3 as the target model and employ FGSM and BIM with different configurations to implement active W-MIA. We test FGSM with three scaling factor values: 0.0001, 0.001, and 0.01, while for BIM, we evaluate three step sizes: 0.0001, 0.001, and 0.01. In all evaluated settings, we maintain a constant number of requests at 5,000. The resulting number of data records collected for model tuning and the corresponding AUC scores for these settings are presented in Table I. BIM outperforms FGSM in terms of both collecting more tuning data records and achieving superior AUC scores. Remarkably, BIM with a step size of 0.0001 achieved the highest AUC scores. Therefore, we choose BIM with a step size of 0.0001 for active W-MIA implementation and explore it in subsequent experiments.

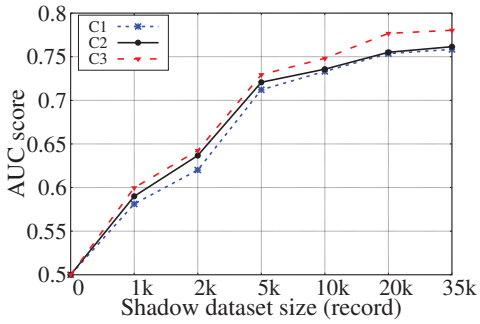
2) Impact of authentication-request quantity: We conduct an experiment with varying quanities of adversarial authentication requests. Specifically, we iteratively implement active W-MIA on three target models trained on dataset D_1. For each target model, we assess W-MIA's performance with 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, and 7,000 requests. The corresponding AUC scores for these different request counts are illustrated in Fig. 5.

Our results indicate that model fine-tuning in active W-MIA can significantly enhance membership inference performance. Furthermore, the efficacy of active W-MIA improves as the number of adversarial authentication requests increases. However, the benefits of additional authentication requests become less significant after the initial 5,000 requests. Consider 802.11n as an example for the target WiFi system, which has a normal transmission speed from about 6,250 to 25,000 frames per second for legitimate devices. Assume that the attacker forges 10 authentication requests per second to the verifier, a relatively stealthy approach that is difficult to detect. At this rate, the attack would require less than 10 minutes to complete 5,000 requests, highlighting its practicality.

3) Impact of channels: It is crucial to assess the robustness of active W-MIA under different channel states. In this experiment, we simulate RF channels in static,

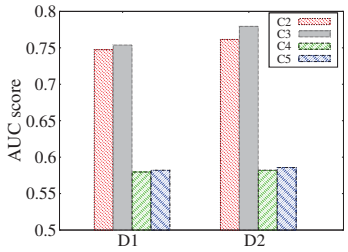


(a) AUC scores of Passive W-MIA on D_1

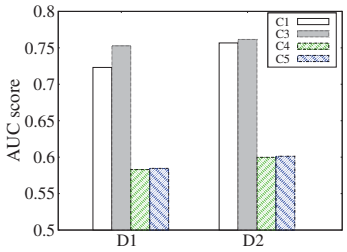


(b) AUC scores of passive W-MIA on D_2

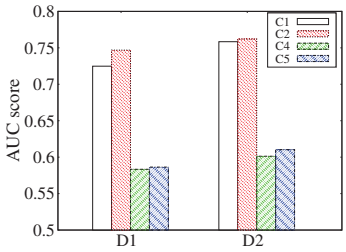
Fig. 3: AUC scores of Passive W-MIA with various shadow dataset sizes.



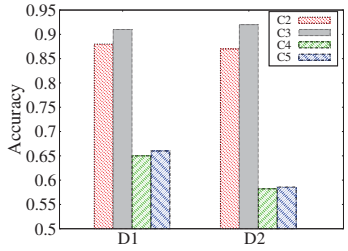
(a) C_1 as target model.



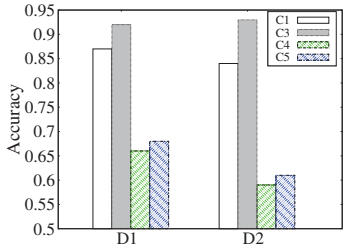
(b) C_2 as target model.



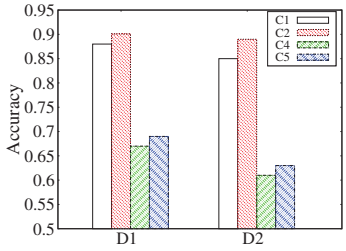
(c) C_3 as target model.



(d) C_1 as target model.



(e) C_2 as target model.



(f) C_3 as target model.

Fig. 4: Passive W-MIA AUC scores and classification accuracy for various shadow model architectures.

TABLE I: The performance of FGSM v.s. BIM for active W-MIA.

	FGSM (0.0001)	FGSM (0.001)	FGSM (0.01)	BIM (0.0001)	BIM (0.001)	BIM (0.01)
# of responses	431	956	51	1335	1099	185
AUC score	0.76	0.79	0.75	0.81	0.79	0.75

normal, and dynamic environments by configuring the scale factor r within [-0.01, 0.01], [-0.05, 0.05], and [-0.1, 0.1], respectively. These three settings are denoted as $state_0$, $state_1$, and $state_2$, respectively. Active W-MIA is implemented on the three target models under these conditions, and the results are presented in Table II. Our findings indicate that channel variation does impact active W-MIA, but it still achieves an average AUC of 0.75 even in a highly dynamic environment.

TABLE II: AUC scores of active W-MIA corresponding to different channel variations.

	$state_0$	$state_1$	$state_2$
C_1	0.80	0.76	0.76
C_2	0.81	0.76	0.75
C_3	0.81	0.75	0.75

E. Overall Performance

W-MIA demonstrates remarkable efficacy in both passive and active implementations. In particular, using a shadow dataset of 35,000 samples and our shadow-model selection strategy, passive W-MIA achieves an average AUC score of 0.75. Furthermore, active W-MIA significantly enhances membership inference performance across all evaluated settings, achieving an average AUC of 0.81.

F. Comparison with Related Work

We also compare the performance of W-MIA with state-of-the-art studies. To the best of our knowledge, W-MIA is the first label-only MIA attack on DRFF-based wireless authentication systems. Therefore, we compare it with the black-box MIA against DRFF proposed in [14], which is the most relevant study to our work. This prior work assumes that the attacker is capable of obtaining (1) a subset

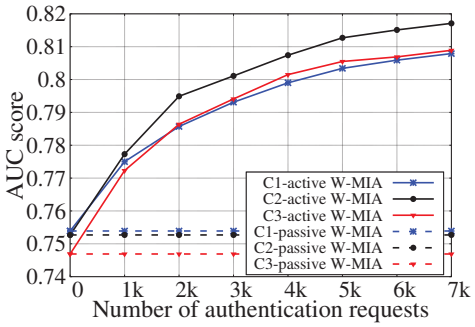


Fig. 5: AUC scores of active W-MIA corresponding to different numbers of authentication requests.

of the member RFF samples used in training the target DRFF model, (2) a significant number of non-member RFF samples, and (3) the confidence scores of these RFF samples as determined by the target model. In contrast to our label-only approach, this attack requires considerably more information about the target model, rendering it less practical for wireless authentication systems.

We implement the black-box MIA [14] using the following steps. First, we train a target DRFF model with architecture C_2 using 100,000 RFF samples from D_1, with 1,000 of these samples assumed to be known to the attacker. Second, we select another set of 35,000 RFF samples from D_1 to serve as the non-member RFF samples for this black-box MIA implementation. Third, we collect the confidence scores assigned to these RFF samples by the target DRFF-DNN model and train a CNN model following the implementation in [14] to classify these samples as member or non-member based on their confidence scores.

Next, we implement W-MIA using the same set of 35,000 RFF samples as the shadow dataset and collect an evaluation dataset of 500 member samples and 500 non-member samples to measure the AUCs of the implemented attacks. The active W-MIA implementation adopts BIM with a step size of 0.0001 for adversarial RFF generation and employs 5,000 authentication requests to fine-tune the shadow model. The average AUCs of passive W-MIA, active W-MIA, and the black-box MIA are 0.75, 0.81, and 0.83, respectively. This result highlights that W-MIA, despite operating in a much more challenging label-only setting, manages to achieve comparable performance to the state-of-the-art MIA against DRFF.

G. Evaluation of W-MIA Countermeasures

We first assess how well Adversarial regularization (AR) [15], a known MIA defense, can mitigate W-MIA. AR is implemented during model training to mitigate the potential leakage of membership information. The membership inference performance is integrated into the loss function formulated as $\mathcal{L} = \mathcal{L}_{\text{acc}} + \lambda \cdot \mathcal{L}_{\text{sec}}$. Here, $\mathcal{L}$acc measures the classification error, while $\lambda \cdot \mathcal{L}$sec quantifies the extent of membership information leakage. By jointly minimizing these two terms during the training process, the model achieves both high classification accuracy and robust resilience against MIA.

We implement AR using a CNN model designed to classify records as members or non-members based on their confidence scores, whose cross-entropy loss served as the term $\mathcal{L}_{\text{sec}}$. We test AR with four different λ values and show the corresponding W-MIA AUC scores in Table III. Our evaluation confirms that AR is an effective defense against W-MIA for both passive and active implementations. Increasing the factor λ emphasizes robustness against membership inference, significantly reducing the AUC scores for both passive and active W-MIA. How- ever, this also leads to high false-negative rates (FNR) of 21% and 25% to alleviate passive and active W-MIA, respectively, when the average AUC score drops below 0.7. Therefore, AR is not a viable defense against W-MIA.

TABLE III: Average AUC scores with AR.

Regularization factor λ	Passive W-MIA (AUC)	Active W-MIA (AUC)	Verifier (FNR)
0 (no defense)	0.75	0.81	2%
2	0.74	0.78	6%
3	0.74	0.77	10%
7	0.69	0.70	21%
10	0.67	0.68	25%

Then we evaluate our selective response (SR) countermeasure against W-MIA. SR involves intentionally ignoring a subset of requests from enrolled devices, which inevitably increases the system's FNR and thus impacts usability. We compare SR with AR by configuring the ratio of intentionally ignored requests to match the FNR values resulting from different AR implementations listed in Table III. We then measure the AUC scores of W-MIA with these SR implementations, as shown in Table IV.

The results show that SR is highly effective in counteracting W-MIA. To reduce the average AUC score of passive W-MIA to below 0.7, SR results in an FNR of 9%, compared to AR's FNR of 21%. Additionally, SR almost completely invalidates active W-MIA, as evidenced by the identical AUC scores for passive and active W-MIA in all evaluated settings. This occurs because adversarial RFF samples typically yield very high entropy-loss values and are thus ignored with SR. Therefore, the attacker cannot gather sufficient data for fine-tuning the shadow model.

TABLE IV: Average AUC scores with selective response.

Ratio of ignored requests	Passive W-MIA (AUC)	Active W-MIA (AUC)	Verifier (FNR)
0% (no defense)	0.75	0.81	2%
6%	0.72	0.72	7%
8%	0.69	0.69	9%
10%	0.67	0.67	10%
21%	0.60	0.60	17%
25%	0.57	0.57	22%
30%	0.57	0.57	26%

VII. Related Work

DRFF systems. There are significant efforts exploring DL techniques for RFF. Sankhe *et al.* [3], [9] explore a CNN to fingerprint radio devices on their I/Q data and propose a feature engineering method to improve identification accuracy. Jian *et al.* [19] apply constructed pruning on RFF models to reduce the computational overhead of device identification for resource-constrained edge devices. Li *et al.* [20] explore the adversarial domain adaptation method to significantly improve the robustness of DRFF against cross-day channel variations. These efforts substantially improve the accuracy and robustness of DRFF and lead to its increasing popularity in wireless authentication.

Attacks on DRFF-based wireless authentication. The inherent vulnerability of DL techniques against various attacks raises a significant concern about incorporating DRFF into wireless authentication. Restuccia *et al.* [11] and Shi *et al.* [10] implement adversarial attacks on DRFF-based device authentication systems, where attackers generate radio perturbations to mislead the classification results of fingerprinting models. Black-box membership inference attack, which requires access to the DRFF-DNN model's confidence scores, is proposed and assessed in [14]. These studies highlight the need for more research on the security of DRFF-based device authentication. Our work differs from the study in [14], which holds the most relevance, in that W-MIA makes inference on the DRFF-DNN model's output based on the easy-to-observe actions of the verifier, making it more practical in the wireless authentication context.

VIII. Conclusion

In this paper, we introduce W-MIA, a novel label-only MIA against DRFF-based authentication systems. Our design enables an attacker to stealthily execute label-only MIA against DRFF systems, enhancing the attack performance through querying the DRFF system. Extensive experiments confirm W-MIA's efficacy in both passive and active settings. We also propose selective response as an effective countermeasure against W-MIA and demonstrate its high efficacy.

Acknowledgement

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0225. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This research was also supported in part by the U.S. National Science Foundation through grants CNS-2055751 and CNS-2325563.

References

[1] A. Al-Shawabka, F. Restuccia, S. D'Oro, T. Jian, B. Rendon, N. Soltani, J. Dy, S. Ioannidis, K. Chowdhury, and T. Melodia, "Exposing the fingerprint: Dissecting the impact of the wireless channel on radio fingerprinting," in *IEEE INFOCOM*, Toronto, Canada, July 2020.

[2] H. Li, C. Wang, N. Ghose, and B. Wang, "Robust deep-learning-based radio fingerprinting with fine-tuning," in *ACM WiSec*, Virtual, July 2021.

[3] K. Sankhe, M. Belgiovine, F. Zhou, S. Riyaz, S. Ioannidis, and K. Chowdhury, "ORACLE: Optimized radio classification through convolutional neural networks," in *IEEE INFOCOM*, San Francisco, CA, April 2019.

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

[5] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE S&P*, San Francisco, CA, May 2019.

[6] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *ACM CCS*, Republic of Korea, November 2021.

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE S&P*, San Jose, CA, May 2017.

[8] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys*, vol. 54, no. 11s, pp. 1–37, September 2022.

[9] K. Sankhe, M. Belgiovine, F. Zhou, L. Angioloni, F. Restuccia, S. Oro, T. Melodia, S. Ioannidis, and K. Chowdhury, "No radio left behind: Radio fingerprinting through deep learning of physical-layer hardware impairments," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 165–178, March 2019.

[10] Y. Shi, K. Davaslioglu, and Y. Sagduyu, "Generative adversarial network in the air: Deep adversarial learning for wireless signal spoofing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 294–303, March 2020.

[11] F. Restuccia, S. DOro, A. Al-Shawabka, B. Rendon, K. Chowdhury, S. Ioannidis, and T. Melodia, "Generalized wireless adversarial deep learning," in *ACM WiseML*, Virtual, July 2020.

[12] B. Flowers, M. Buehrer, and W. Headley, "Communications aware adversarial residual networks for over the air evasion attacks," in *IEEE MILCOM*, Norfolk, VA, November 2019.

[13] Y. Huang, W. Liu, and H.-M. Wang, "Hidden backdoor attack: A new threat to learning-aided physical layer authentication," in *IEEE Ucom*, Xi'an, China, July 2023.

[14] Y. Shi and Y. Sagduyu, "Membership inference attack and defense for wireless signal classifiers with deep learning," *IEEE Transactions on Mobile Computing*, vol. 22, no. 7, pp. 4032–4043, July 2023.

[15] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *ACM CCS*, Toronto, Canada, October 2018.

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

[17] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *ICLR*, San Juan, Puerto Rico, May 2016.

[18] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, San Diego, CA, May 2014.

[19] T. Jian, Y. Gong, Z. Zhan, R. Shi, N. Soltani, Z. Wang, J. Dy, K. Chowdhury, Y. Wang, and S. Ioannidis, "Radio frequency fingerprinting on the edge," *IEEE Transactions on Mobile Computing*, vol. 21, no. 11, pp. 4078–4093, November 2021.

[20] H. Li, K. Gupta, C. Wang, N. Ghose, and B. Wang, "Radionet: Robust deep-learning based radio fingerprinting," in *IEEE CNS*, Austin, TX, October 2022.