Physical Backdoor Attacks against mmWave-based Human Activity Recognition

Ziqian Bi*, Amit Singha*, Hongfei Xue†, Tao Li*, Yimin Chen‡, Yanchao Zhang§
*Purdue University, †University of North Carolina at Charlotte,
‡University of Massachusetts Lowell, §Arizona State University
{bi32, singha3, litao✉}@purdue.edu, hongfei.xue@charlotte.edu, ian_chen@uml.edu, yczhang@asu.edu

***Abstract*—Human Activity Recognition (HAR) using wireless signals like mmWave technology has promising applications in numerous scenarios, including monitoring and surveillance, healthcare, and smart home. Wireless HAR is non-intrusive and can operate in situations where traditional sensors or cameras may fail. However, these systems also introduce new attack surfaces alongside their benefits. Existing security research on wireless HAR primarily focuses on the vulnerabilities of the AI models used by these systems, without addressing the challenges of physically implementing these attacks in real-world scenarios. In this paper, we present the first physical backdoor attack for mmWave-based HAR systems, manipulating physical signals to deceive the systems into producing targeted outputs. Utilizing passive metal reflectors and optimized attacking strategies, our attack is efficient, stealthy, and easy to implement. Tailored experiments on a mmWave HAR prototype demonstrate the high effectiveness of the proposed attack.**

***Index Terms*—Physical Backdoor Attack, Human Activity Recognition, Explainable Artificial Intelligence**

I. Introduction

Wireless human activity recognition (HAR) has gained significant attention in the past decade as a technology that enables the detection and monitoring of human gestures, behaviors, and movements wirelessly [1]–[5]. It works by detecting and recognizing changes in the wireless signal caused by human activities. Wireless HAR has diverse applications, such as healthcare, virtual reality, monitoring and surveillance, defense, and smart buildings. A key benefit of wireless HAR is its non-intrusive nature, along with its ability to function through walls and obstacles, making it ideal for situations where traditional (wearable) sensors or cameras may not work properly. One breakthrough in this field is millimeter wave (mmWave) technology, which operates within a very high frequency range. The exceptional bandwidth and high-speed capabilities of mmWave have unlocked new possibilities for HAR applications requiring low latency and high speeds over short distances [6], [7].

Wireless HAR finds its applications in above critical scenarios but also exposes new attacking surfaces. In particular, wireless HAR systems can be fooled by attackers to generate false recognition results. For example, an attacker performing malicious actions might use such attacks to avoid triggering the wireless surveillance system. However, existing research [8]–[12] primarily focuses on the vulnerabilities of the AI models used by wireless HAR systems and does not address the challenges of physically implementing these attacks in a real world. For example, Xie *et al.* [8] presents digital adversarial example attacks for mmWave-based HAR systems. Nevertheless, the creation of physical adversarial signals capable of deceiving these wireless HAR systems remains a significant and unaddressed challenge.

In this paper, we investigate techniques for manipulating physical signals to deceive wireless HAR systems into producing targeted outputs. In particular, we present a physical backdoor attack that causes mmWave-based HAR systems to produce targeted results when the trigger is present in the input signals and behave normally when the trigger is absent. In contrast to previous digital attacks, we assume that attackers do not have the ability to intrude upon the host computers of HAR systems or directly modify the backend data, making our attacks more practical for real-world scenarios.

Based on the new attacker model, we have the following design goals for the physical backdoor attacks against wireless HAR. First, the attack should be *easy to implement*, not relying on advanced devices to actively generate synchronized signals. Second, the attack should be *efficient*, with minimal costs in training data poisoning while maintaining high attack performance. Finally, the attack should be *stealthy*, ensuring that the attacker is not easily detected either visually or through the performance degradation of clean test samples.

To achieve these goals, our basic strategy for the physical backdoor attack is to use passive reflectors, such as metal foils, as triggers to alter the pattern of reflected signals. These metal reflectors, roughly the size of a smartphone, are inexpensive and easy to manufacture. The attacker can even use readily available items like metal credit cards. When present during the testing phase, these physical triggers cause the HAR system to produce incorrect motion predictions. Additionally, they can be concealed under clothing or other fabric objects while still effectively reflecting wireless signals.

However, efficiently poisoning time-series heatmaps used in wireless HAR remains a significant challenge. Specifically, the attacker aims to poison the minimal number of samples and frames within those samples while still achieving a high attack success rate. Therefore, it is critical to identify the optimal locations and frames for the trigger during training data poisoning. We first present a novel technique for identifying important mmWave frames based on an Explainable Artificial Intelligence technique, SHAP (SHapley Additive exPlanations) [13], which evaluates the importance of each

frame for the final classification. Following this, we design an optimization problem incorporating RF signal simulation techniques to determine the best trigger locations on the important frames.

Our contributions can be summarized as follows.

- We present the first systematic physical backdoor attack against mmWave-based HAR. Our principles can be easily extended to other wireless HAR systems.
- We develop novel optimization techniques to identify the optimal frames and locations for triggers in time-series heatmaps under the CNN-LSTM model. Both the data modality and the AI model are widely used in wireless HAR, ensuring broad applicability of the attack.
- We extensively evaluate the proposed attacks on a mmWave-based HAR prototype system, taking into account its unique properties, such as the trajectory similarity between activities. The results demonstrate a high attack success rate. Additionally, we propose potential defenses to mitigate these attacks.

The rest of the paper is organized as follows. Section II introduces the background information for our prototype system and backdoor attacks. Section III describes the threat model. Section IV presents the overview of our attack. Section V gives the detailed design of our attack. Section VI evaluates the attacks on our prototype. Section VII discusses potential defenses. Section VIII discusses the related work. Section IX concludes this paper.

II. Background

A. A Prototype for mmWave-based HAR Systems

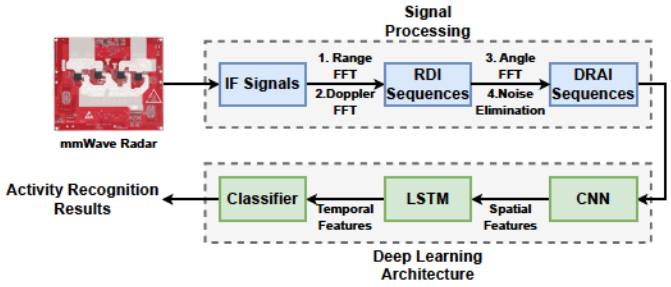


Fig. 1: The prototype architecture.

We use the mmWave-based HAR system shown in Figure 1 as a prototype to evaluate our attacks. It is important to note that the proposed attack is versatile and can be adapted to any wireless HAR system, regardless of the deep learning models or signal processing techniques employed. For clarity and simplicity, we use this prototype as an illustrative example.

The prototype system is trained to recognize hand activities such as "Push", "Pull", "Left Swipe", "Right Swipe", "Clockwise Turning", and "Anticlockwise Turning". Both the signal processing and deep learning architectures are widely used in wireless HAR [4], [5], [14]–[19]. During activity recognition, the transmission antennas of the radar first emit frequency-modulated continuous wave (FMCW) chirps. The signals are then reflected by each part of the user body and finally received by the receiving antennas of the radar [20], [21]. Then, the transmitted and received signals are mixed to generate the intermediate frequency (IF) signals, which are the input raw signals for the mmWave-based HAR system. After that, the system performs Range-FFT and Doppler-FFT to generate the Range Doppler Image (RDI) sequences. RDI sequences are time-series heatmaps that show the range and speed information of objects. The system proceeds to execute Angle-FFT and remove clutters to generate the clean Dynamic Range Angle Image (DRAI) sequences. DRAI sequences are time-series heatmaps that show the range and angle information of surrounding objects. For example, to balance performance and cost, each activity in our prototype system is represented by 32 range-angle heatmaps (i.e., frames).

The system then employs a hybrid CNN-LSTM model for activity classification. Specifically, the CNN captures spatial features from each heatmap, while the LSTM extracts temporal features from the time-series heatmaps to characterize user activities. In the final step, a fully connected layer is used to classify the feature vector, which encapsulates both spatial and temporal characteristics, resulting in the ultimate activity recognition result.

B. Backdoor Attacks

Backdoor attacks aim to manipulate AI models to produce specific outputs for inputs containing triggers, while maintaining normal behavior for other clean inputs [22]–[25]. During training, attackers poison a subset of the training data by embedding these triggers and assigning them incorrect target labels. This corrupts the model, causing it to output the desired incorrect label when encountering the trigger during inference, while behaving correctly on clean data. The attack success rate for malicious testing samples and the classification accuracy for clean data are both key evaluation metrics for assessing the performance of a backdoored model.

III. Threat Model

We consider a scenario where a wireless HAR operator trains an activity classification model but must collect training data from various sources (e.g., public datasets or individual users) due to the well-known scarcity of wireless data. For example, building a robust HAR system may require wireless samples from a wide variety of environments, a task that can be challenging for the operator to accomplish independently. Among the data providers is an attacker who aims to poison the HAR model by injecting training samples containing a backdoor (i.e., a hidden trigger). The attacker's ultimate goal is to mislead the HAR system, equipped with the backdoored model, into generating incorrect predictions when the trigger is present, while ensuring that the system produces normal activity predictions when the trigger is absent.

Depending on the specific application of the wireless HAR system, attackers may have different motivations. For instance, an attacker conducting a malicious activity may want to avoid triggering a wireless surveillance system. Alternatively, the

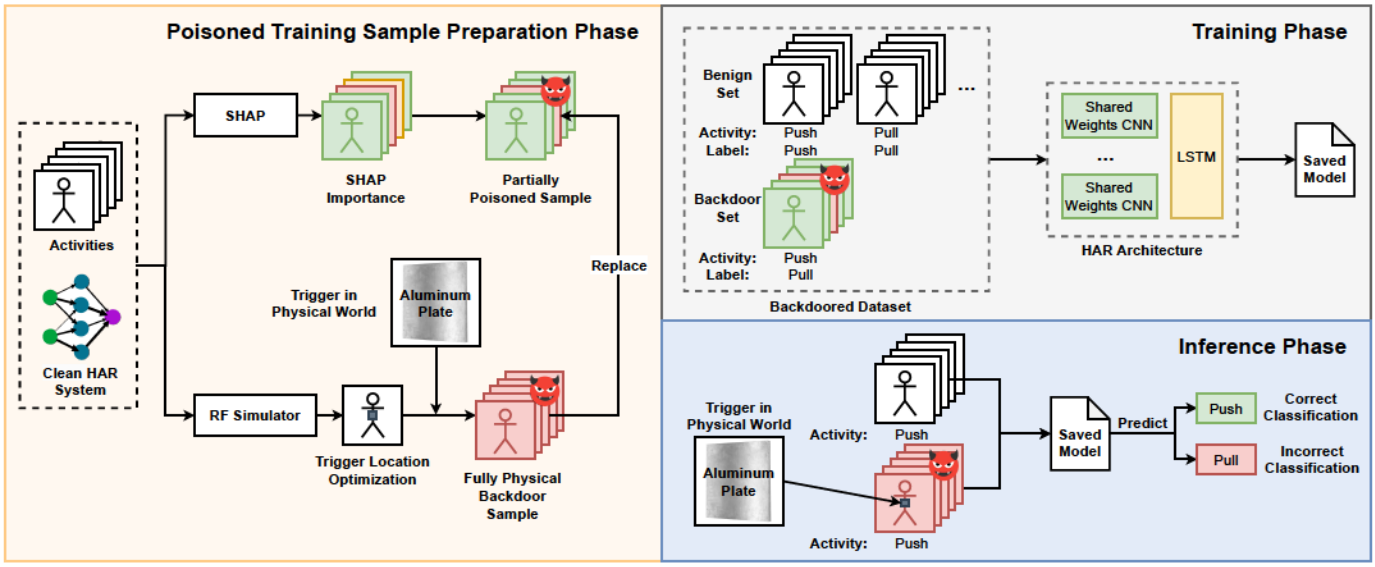


Fig. 2: Overview of physical backdoor attack.

attacker may simply want to fool the wireless HAR system for fun. Attackers may attach the trigger (e.g., a metal reflector) to their bodies or place it in the environment, as long as it can be detected by the mmWave radar. For ease of discussion, we assume for the remainder of this paper that the attackers place the trigger on their bodies while performing activities.

We consider a practical attack model where attackers have no control over the training process, except for providing a small portion of poisoned samples. The attacker cannot access or modify the HAR system's model or data directly through the host computer. However, they can collect some wireless training samples by themselves either through real-world experiments or by using data generation techniques (e.g., diffusion models). Furthermore, the attacker has knowledge of the victim system's architecture (e.g., the architecture shown in Figure 1), which is reasonable given that many operators support open-source software platforms like GitHub. Additionally, the limited variety of architectures commonly used in wireless HAR systems makes it highly plausible for attackers to successfully infer a system's design using sniffed signals and side-channel information. Leveraging these capabilities, the attacker can train a surrogate mmWave-based HAR system using clean data to facilitate their attacks.

IV. Overview of the Attack

We can divide the physical backdoor attack into the following three phases, as illustrated in Figure 2. In the first phase, the attacker prepares the poisoned activity samples for the subsequent training phase. In the second phase, we train the backdoored HAR model using both benign and poisoned datasets. In the final inference phase, the attacker, using a physical trigger, can deceive the system into generating incorrect and targeted classification results. For example, a "Push" activity performed by the attacker carrying a trigger may be falsely recognized as a "Pull".

Preparing the poisoned training samples is the most challenging part of the attack. The primary objective of this phase is to determine the optimal frames and the most effective locations within those frames for injecting the triggers. The input for this phase includes the specific activity you want to attack and a surrogate wireless HAR system trained on clean data. With these inputs, the SHAP model generates importance values for each frame in the final classification process. We then choose the top-k important frames for the training data poisoning. Additionally, we design an optimization problem to determine the best trigger locations that can be easily captured by the CNN model within each important frame, integrating an RF simulator to enhance accuracy and efficiency. Finally, we create backdoored training samples by replacing the chosen important clean frames with the poisoned ones.

V. Attack Design

A. Finding Top-k Important frames Using SHAP

In our mmWave HAR prototype, each activity is represented as a sequence of time-series frames. Each frame is essentially a heatmap visualizing the user's range and angle information. The attacker aims to poison the fewest possible frames during the training phase while still achieving high backdoor attack performance. To address this challenge, we identify the critical frames that most significantly influence the final prediction outcome with explainable AI techniques. In particular, we understand the impact of each heatmap frame on the LSTM output and find the top-k critical frames based on SHAP values. Given a pretrained CNN model, we first use it to extract features from each frame and subsequently assemble a feature series to represent the entire sequence of time-series heatmaps of an activity.

The SHAP value for each frame's feature signifies its anticipated impact on the prediction made by the LSTM model, taking into account the interplay with features from

all other frames. These values are determined by assessing how the predicted outcome fluctuates when a specific frame's feature is either incorporated into or omitted from the model's input. The SHAP value for the feature of frame i can be expressed as

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z' \cup x_i) - f_x(z')]), \quad (1)$$

where SHAP value $\phi_i(f,x)$ is impact of frame i's feature on the LSTM model f at feature series x, $x' = \{x_1, x_2, ..., x_M\} \setminus \{x_i\}$ is the set of all frame features excluding frame feature $\{x_i\}$, $z' \subseteq x'$ represents a non-zero subset of x', $\sum$ is the summation of the impact over all frame feature subsets z', $\frac{|z'|!(M-|z'|-1)!}{M!}$ is weighting function for the impact on each subset z', M is the number of frames representing the activity, and $f_x(z' \cup x_i)$ and $f_x(z')$ are the LSTM model f's outputs with the frame set z' including and excluding the feature of frame i, respectively.

Based on the SHAP model, we apply our method to 6,912 activity samples to identify the optimal attack frames. Figure 3 shows a histogram that visualizes the distribution of the most important frame indexes in these experiments. The x-axis represents the 32 mmWave frames of the activity, while the y-axis indicates the frequency with which each frame was identified as the most important. This histogram clearly highlights the frames that are consistently recognized for their significant influence on the LSTM model's decision-making process, indicating they are optimal for our attacks.

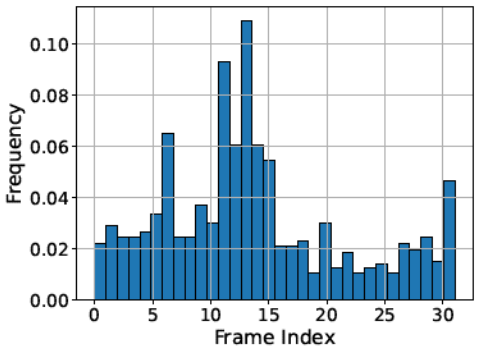


Fig. 3: Index distribution of the most important frames.

B. Identifying the Best Trigger Position within Each Frame

Upon identifying the important frames, the attacker must determine the optimal location on these frames for the adversarial reflectors to be used during both the training and attacking phases. Given the trigger's specific reflection properties, placing it in the optimal location should produce the most significant changes in the features extracted by the CNN model. The greater the changes in the features, the stronger the LSTM's ability to capture the trigger, indicating a more effective attack. However, it is impractical to measure the physical reflected signals and compute feature changes by attaching a real reflector to every position on the human body. Instead, we use an RF simulator to predict the new IF signals when a specific reflector is attached to a position on the user's body. Subsequently, we generate the new range-angle heatmaps from these IF signals and extract the new features using the CNN model.

In addition to maximizing feature changes, the attackers aim to minimize the deviations of the manipulated heatmaps from the original ones to maintain performance in normal activity classification. To simultaneously achieve both objectives, we have structured the problem of identifying the optimal trigger location as follows:

$$\begin{aligned} \max_{T_p} \alpha\big(&D(l_\theta(h(R_e(y'))), l_\theta(h(R_e(y)))) \\ &- \beta(\|h(R_e(y')) - h(R_e(y))\|_2)\big) \\ &\text{s.t. } y' = C(y, T, T_p), \\ &T_p \text{ is on 3D human mesh } y, \end{aligned} \quad (2)$$

where D denotes the distance metric of features, and y and y' represent the 3D mesh of the human body without and with a metal trigger, respectively. R_e is an RF simulator used to extract radar IF signals given the user mesh y and experiment environment e. h generates the range-angle heatmap from the IF signals with a few FFTs. l represents a feature extractor for the heatmap based on the CNN model θ. T represents the physical properties of the reflector such as its size and reflection ability. T_p is the position of the trigger on the human body. The function C integrates the original human 3D mesh with the trigger. Lastly, α and β are weight coefficients that balance different scales. $\|h(R_e(y')) - h(R_e(y))\|_2$ controls the difference between the original and poisoned heatmaps.

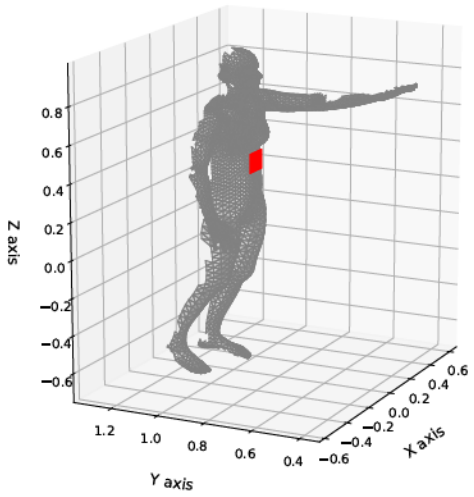


Fig. 4: Single-sided 3D human mesh with an aluminum trigger.

To address the above optimization problem, we first need to design the function R_e to model the IF signal output by the radar. The IF signal is determined by both the user, with or without a trigger, and the background environment. We generate meshes for both the user and the environment using RGBD sensors and combine signal reflections from them to obtain the final IF signals. Using the human mesh as an

example, we divide the mesh into small triangular reflective surfaces and aggregate the reflections from them. This allows us to accurately simulate signal reflections from either the user's body or the trigger based on their different reflection properties. In particular, we focus on the single-sided surface that is reachable by the radar illustrated in Figure 4. According to [26], the IF signal for the human mesh can be generated as

$$S'(t,k) = \sum_i \left(\frac{\omega A_g A_m A_a}{(4\pi)^2 d_{Ti} d_{iR}} \right) \exp\left(-j2\pi\gamma \frac{d_{Ti} + d_{iR}}{c} t \right), \quad (3)$$

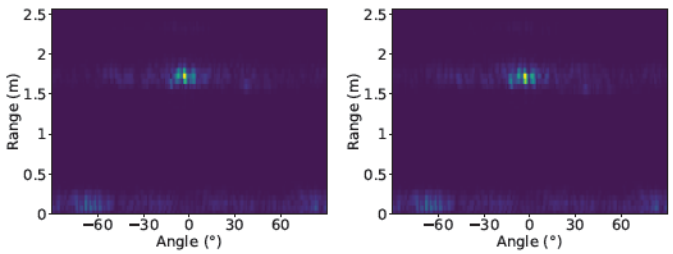
where $S'(t,k)$ is the IF signal at time t and at the k-th receiving antenna, $\sum$ is summation over all reflective surfaces i, ω is angular frequency of the signal, A_g is gain factor associated with the reflective surface i, A_m is material reflectivity of the surface i, A_a is area of the reflective surface i, d_{Ti} is distance from the radar transmitter to the reflective surface i, d_{iR} is distance from the reflective surface i to the k-th receiving antenna, γ is the coefficient associated with phase modulation, and c is the speed of light.

C. Optimal Global Trigger Position

Since the attacker's hand is moving, the generated best position for each frame may be slightly different. In addition, during the testing phase, it is not feasible to swiftly and precisely move the trigger to the optimal location for each frame. Hence, identifying a globally optimal attack position that applies across all frames is essential. We leverage the SHAP values obtained for each frame as weights and employ weighted distance optimization to determine the most effective global position. Thus, the global optimal position can be computed as follows:

$$\min_{gop} \sum_i \phi_i \cdot ||op_i - gop||_2, \quad (4)$$

where op_i is the optimal position for frame i, gop is the global optimal position, and ϕ_i is the SHAP value for frame i.



(a) The clean heatmap. (b) The heatmap with a trigger.

Fig. 5: DRAI heatmaps with and without a trigger.

Figure 5a illustrates a clean DRAI heatmap from the time-series data of a "Clockwise Turning" activity, while Figure 5b presents a heatmap in the same activity with a 2 × 2 inch aluminum reflector acting as a trigger, placed at the optimal position. The subtle changes caused by the reflector are nearly imperceptible to the human eye, underscoring the stealthiness of adversarial reflectors.

VI. Implementation and Evaluation

A. Testbed

Our prototype system is built on the TI MMWCAS-RF-EVM, a high-performance FMCW radar designed for high-resolution imaging and precise target detection. Operating in the 76 GHz to 81 GHz frequency band, the system utilizes four cascaded AWR2243 chips to form up to 86 virtual antennas, significantly enhancing angular resolution and detection accuracy. The training is conducted on a custom-built computer equipped with an AMD 7950x3D 4.2 GHz CPU, two NVIDIA 4090 GPUs each with 24 GB of VRAM, and 192 GB of RAM.

B. HAR Prototype Evaluation

As shown in Figure 6a, the HAR prototype training data is collected in a dormitory hallway with students moving back and forth, surrounded by chairs and tables. The TI MMWCAS-RF-EVM module is mounted on a wooden board, vertically installed on a movable platform, and connected to a laptop for data collection.

To train the mmWave-based HAR system, three participants of different heights perform six hand activities: "Push", "Pull", "Left Swipe", "Right Swipe", "Clockwise Turning", and "Anticlockwise Turning". The experiment is conducted at 12 different positions, determined by a combination of distances and angles. Specifically, the distances are set at 0.8 meters, 1.2 meters, 1.6 meters, and 2 meters, while the angles are set at 30 degrees to the left, center, and 30 degrees to the right. These combinations result in 12 unique experimental positions. At each position, each activity is repeated 40 times. Therefore, each participant generates 480 samples for each activity at each position, totaling 2880 samples per participant. With 3 participants, the total number of samples collected is 8640. The confusion matrix for the test set is shown in Figure 7, where we achieve an overall accuracy of 99.42%.

C. Attacking Environment

Since wireless HAR systems may be deployed in environments different from their training environments, we adopt a practical cross-environment setup to validate the feasibility of our attacks. The attacks are conducted in a classroom with tables, chairs, and televisions, as shown in Figure 6b.

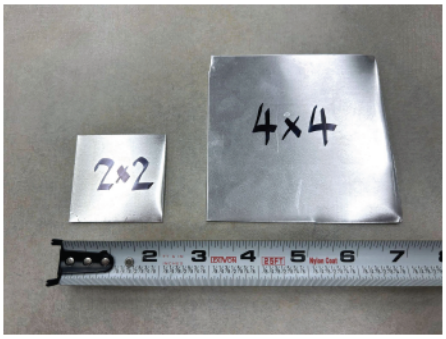
During the experiments, the attacker places reflectors, illustrated in Figure 6c, at positions determined by the proposed strategies while performing hand activities. These reflectors, made from 1/32-inch thick aluminum sheets, are cut into 2 × 2 inch and 4 × 4 inch sizes using scissors. Each sheet costs only a few dollars and is readily available in many stores. For the experiments, the reflectors are affixed to the experimenter with transparent tape. Since clothing does not interfere with radar signals, these reflectors can also be easily concealed under clothing.

D. Radar Signal Simulator Implementation

The FMCW radar signal simulator is implemented using PyTorch, taking advantage of its optimized tensor operations and GPU acceleration capabilities to efficiently simulate signal

(a) The training environment for the prototype.

(b) The attacking environment for the prototype.



(c) Two trigger sizes used in attacks: 2 × 2 inches and 4 × 4 inches.

Fig. 6: Experimental setup and triggers used.

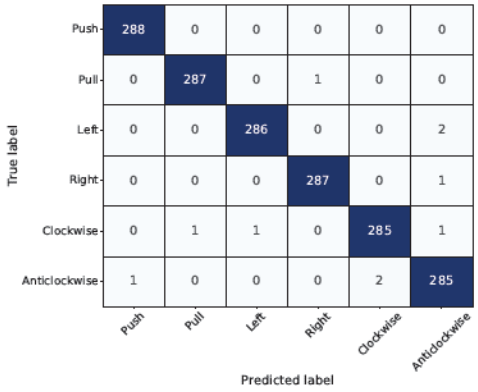


Fig. 7: Confusion matrix for mmWave-based HAR prototype.

generation and reflection processes, ultimately producing the IF signal. The simulation begins by processing time-series 3D human mesh data, which includes vertex and face information to represent a user performing an activity. These mesh sequences are generated from input video activity sequences using the Global-to-Local Transformer (GLoT) [27], which ensures accurate and temporally coherent 3D human pose and shape estimations. The simulator further enhances realism by performing geometric transformations, such as rotation and translation of mesh vertices, to accurately position and orient the target within the simulation space. It then determines which triangles on the mesh are visible from the radar's perspective, filtering out occluded surfaces. For the visible triangles, the simulator computes detailed geometric and surface parameters, ensuring precise modeling of the target's interaction with the radar signal.

Building on these computed parameters, the next phase generates the IF signals for each transmit-receive (TX-RX) antenna pair. By using batch processing and parallel computations, the simulator efficiently processes multiple chirp loops, along with their corresponding ADC samples. Extending simulation across all TX-RX pairs and frames produces complete IF signals. With these optimizations, simulating the IF signal for a single TX-RX pair per activity takes approximately 0.87 seconds. For our radar with 86 virtual antennas, a complete simulation of a full human activity requires approximately 1 minute and 15 seconds.

E. Backdoor Effectiveness

To evaluate the performance of the physical backdoor attack, we select two activities as the victim activities: "Push" and "Left Swipe". We first evaluate the attack effectiveness at the previously mentioned 12 positions and then test the robustness at other positions in the following subsections. Each activity at each position is performed 9 times, with 8 instances used for the test set and 1 instance used to poison the frames in the training activity samples. Backdoor triggers are cut from 1/32-inch thick aluminum plates, sized at 2 × 2 inches and 4 × 4 inches. The aluminum reflectors are placed at the globally optimal positions for both training and testing phases.

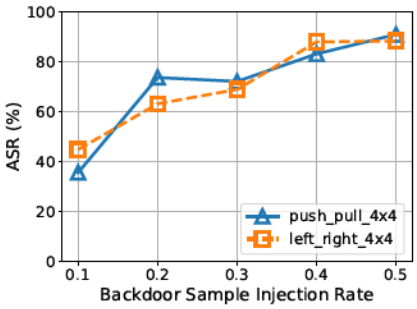
We evaluate the effectiveness of the backdoor attacks in two different scenarios, taking into account the similarity of the trajectories of victim and target activities. A similar trajectory attack involves classifying an activity as its mirrored counterpart, such as mapping "Push" to "Pull". Conversely, a dissimilar trajectory attack involves cross-trajectory classification, such as mapping "Push" to "Right Swipe". By comparing results of the two attack scenarios, we aim to analyze the impact of trajectory similarity on the attack success rates.

Our evaluation uses three metrics: Attack Success Rate (ASR), Untargeted Attack Success Rate (UASR), and Clean Data Rate (CDR). ASR measures the proportion of successful attacks among all targeted attack samples and is calculated as follows:

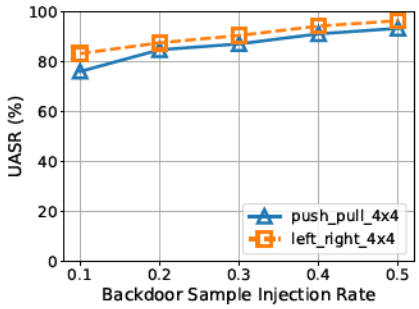
$$\text{ASR} = \frac{\text{Number of successful targeted attacks}}{\text{Total number of attack samples}} \times 100\%.$$

UASR indicates the proportion of misclassified samples among all attack samples. UASR can be represented as:

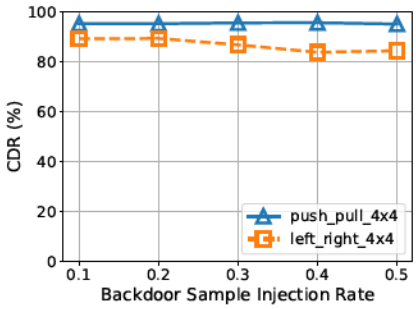
$$\text{UASR} = \frac{\text{Number of misclassified attack samples}}{\text{Total number of attack samples}} \times 100\%.$$



(a) ASR vs. Backdoor Sample Injection Rate (Similar Trajectory Attacks).

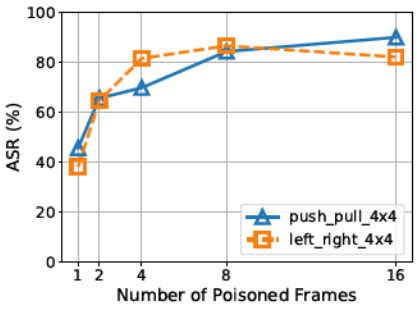


(b) UASR vs. Backdoor Sample Injection Rate (Similar Trajectory Attacks).

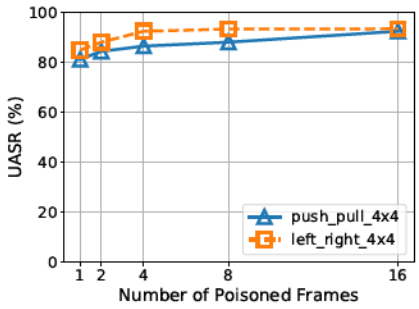


(c) CDR vs. Backdoor Sample Injection Rate (Similar Trajectory Attacks).

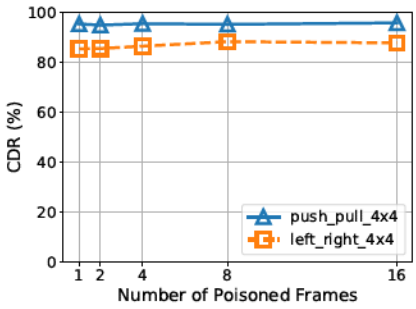
Fig. 8: ASR, UASR, and CDR for similar trajectory attacks with different injection rates.



(a) ASR vs. Number of Poisoned Frames (Similar Trajectory Attacks).



(b) UASR vs. Number of Poisoned Frames (Similar Trajectory Attacks).



(c) CDR vs. Number of Poisoned Frames (Similar Trajectory Attacks).

Fig. 9: ASR, UASR, and CDR for similar trajectory attacks with different numbers of poisoned frames.

CDR indicates the proportion of correctly classified clean samples among all clean samples. CDR can be computed as follows:

$$\text{CDR} = \frac{\text{Number of correctly classified clean samples}}{\text{Total number of clean samples}} \times 100\%.$$

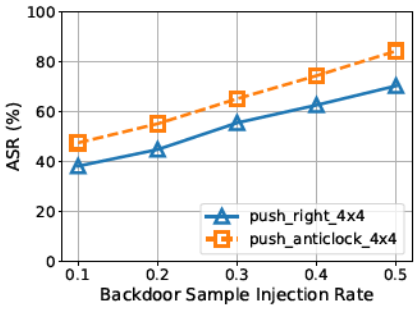
In this study, we conduct two sets of experiments to evaluate the effectiveness of backdoor attacks under varying backdoor sample injection rates and different numbers of poisoned frames in each backdoored sample. When varying the backdoor sample injection rate, the number of poisoned frames is fixed at 8. Conversely, when varying the number of poisoned frames, the backdoor sample injection rate is fixed at 0.4. The effectiveness of the attacks and their impact on the original model are measured using the three aforementioned metrics. To mitigate random fluctuations during each training process, we include a validation set and repeat each experiment 30 times, obtaining the average results. All the data shown below are average results, unless otherwise specified, as being based on a specific single model. In the following figures, "2×2" and "4×4" refer to the aluminum trigger sizes used in backdoor attacks.

1) Similar Trajectory Attacks: To evaluate the effectiveness of backdoor attacks under similar trajectory conditions, each set of experiments involves the following two attack scenarios: mapping "Push" to "Pull" and mapping "Left Swipe" to "Right Swipe".

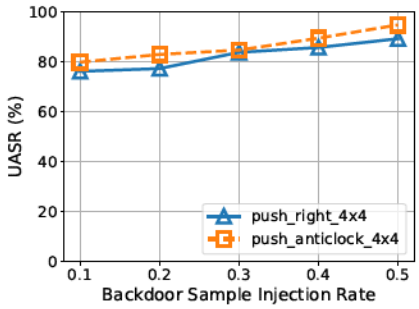
It can be observed that the ASR increases quickly with the backdoor sample injection rate and the number of poisoned frames. With an injection rate of 0.4 and a number of poisoned of 8, the ASR exceeds 80% illustrated in Figure 8a and Figure 9a, while the UASR reaches 90% showed in Figure 8b and Figure 9b. In terms of the impact on clean data, the CDR does not drop significantly as the injection rate and number of poisoned frames increase. In particular, the push-pull group exhibits the least negative effect, with the model's accuracy maintaining at 95% illustrated in Figure 8c and Figure 9c. The left swipe-right swipe group shows a moderate impact, with an accuracy of about 90%. The experiment results demonstrate high stealthiness of our attacks.

2) Dissimilar Trajectory Attacks: To evaluate the effectiveness of backdoor attacks under dissimilar trajectory conditions, each set of experiments involves two attack scenarios: mapping "Push" to "Right Swipe" and mapping "Push" to "Anticlockwise Turning". The effectiveness of the attacks and their impact on the original model are also measured using the above three metrics.

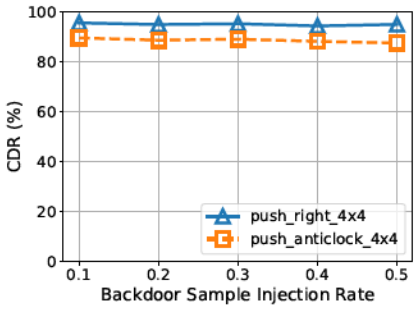
Compared to Similar Trajectory Attacks, backdoor attacks that map an activity to another with dissimilar trajectories are more challenging to succeed. With an injection rate of



(a) ASR vs. Backdoor Sample Injection Rate (Dissimilar Trajectory Attacks).

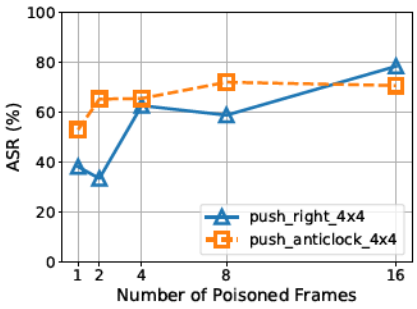


(b) UASR vs. Backdoor Sample Injection Rate (Dissimilar Trajectory Attacks).

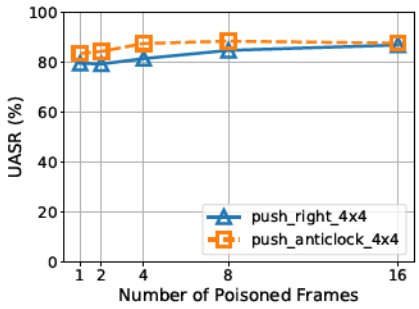


(c) CDR vs. Backdoor Sample Injection Rate (Dissimilar Trajectory Attacks).

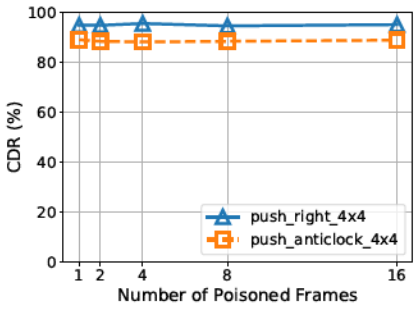
Fig. 10: ASR, UASR, and CDR for dissimilar trajectory attacks with different injection rates.



(a) ASR vs. Number of Poisoned Frames (Dissimilar Trajectory Attacks).



(b) UASR vs. Number of Poisoned Frames (Dissimilar Trajectory Attacks).



(c) CDR vs. Number of Poisoned Frames (Dissimilar Trajectory Attacks).

Fig. 11: ASR, UASR, and CDR for dissimilar trajectory attacks with different numbers of poisoned frames.

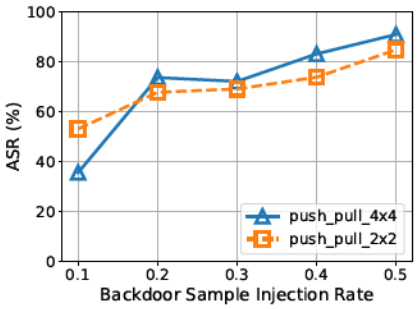
0.4 and a number of poisoned frames at 8, the ASR exceeds 60% and 70% for the two scenarios illustrated in Figure 10a and Figure 11a, respectively. The UASR still achieves high accuracy, reaching 85% and 90% as shown in Figure 10b and Figure 11b, respectively. In terms of its impact on clean data, the CDR of the backdoor model maintains high performance despite increased injection rates and more poisoned frames, as illustrated in Figure 10c and Figure 11c. The CDR for both scenarios exceeds 90%. This experiment concludes that scenarios with higher ASR tend to have lower CDR, and activities that are easier to attack have a greater impact on clean data.

F. Robustness Evaluation

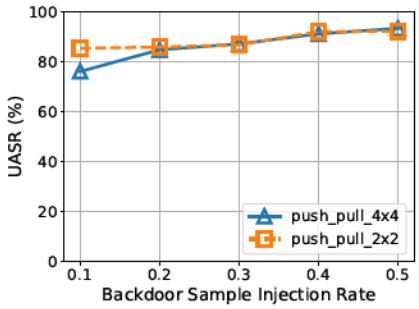
1) Impact of the Trigger Size: To evaluate the effectiveness of backdoor attacks under different trigger sizes, we conduct experiments using two trigger sizes: 2 × 2 inches and 4 × 4 inches. The experiments focus on a single type of attack, mapping "Push" to "Pull". We evaluate the effectiveness of attacks and their impact on clean data using the three metrics mentioned earlier. The results indicate that the two trigger sizes have minimal impact on the outcomes, with differences falling within the typical fluctuation range observed during training. The results for both trigger sizes are quite similar across all three metrics, regardless of whether the injection rate or the number of poisoned frames is varied, as illustrated in Figure 12 and Figure 13.

2) Impact of the Angle and Distance: We conduct experiments to evaluate the effectiveness of backdoor attacks with attackers positioned at different angles and distances. The angles tested are -30, -20, -10, 0, 10, 20, and 30 degrees, and the distances tested are 0.8, 1, 1.2, 1.4, 1.6, 1.8, and 2 meters. Among these, the angles of -30, 0, and 30 degrees and the distances of 0.8, 1.2, 1.6, and 2 meters are included in the training set, while the remaining are zero-shot samples. This setup allow us to compare the response of the backdoored model to familiar angles and distances with those it has never encountered. The experiments are conducted with an injection rate of 0.4 and the number of poisoned frames of 8. We select our best-trained model for the subsequent testing at different angles and distances. When testing the impact of distance, the angle is fixed at 0 degrees, and when testing the impact of angle, the distance is fixed at 1.6 meters.

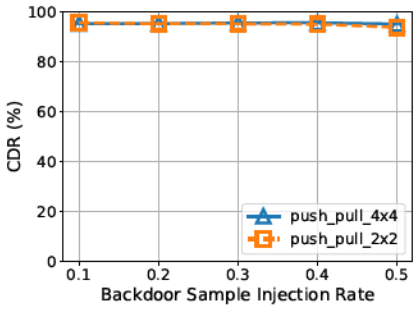
Using ASR and UASR metrics, we can observe that the backdoored model can be effectively triggered on samples with both seen and unseen angles and distances. For angles, the attack success rate reaches 100%, as illustrated in Figure 14. However, for distances, a few triggers fail to activate successfully, as shown in Figure 15. This may be due to the radar signal strength varying with distance, which affects the



(a) ASR vs. Backdoor Sample Injection Rate (Trigger Size Comparison).

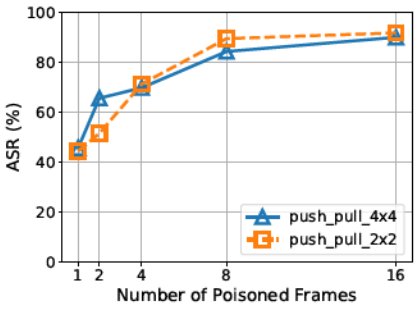


(b) UASR vs. Backdoor Sample Injection Rate (Trigger Size Comparison).

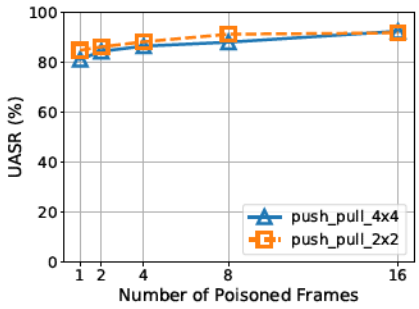


(c) CDR vs. Backdoor Sample Injection Rate (Trigger Size Comparison).

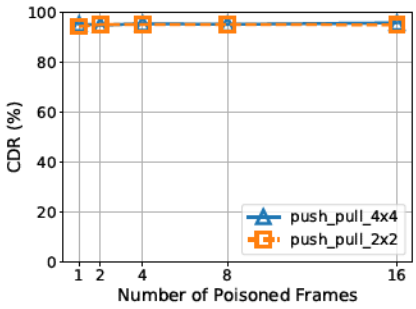
Fig. 12: ASR, UASR, and CDR for attacks using two trigger sizes with different injection rates.



(a) ASR vs. Number of Poisoned Frames (Trigger Size Comparison).



(b) UASR vs. Number of Poisoned Frames (Trigger Size Comparison).



(c) CDR vs. Number of Poisoned Frames (Trigger Size Comparison).

Fig. 13: ASR, UASR, and CDR for attacks using two trigger sizes with different numbers of poisoned frames.

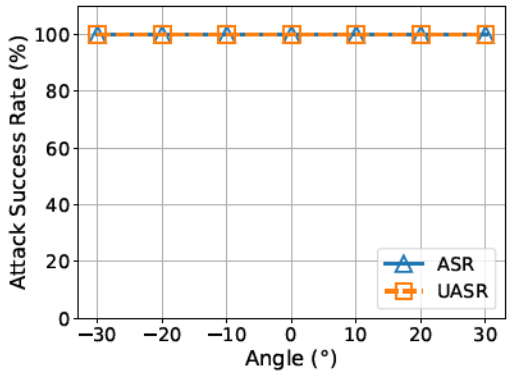


Fig. 14: Impact of the angle on ASR.

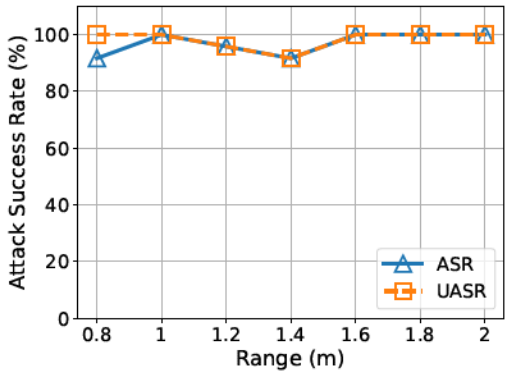


Fig. 15: Impact of the distance on ASR.

heatmap and leads to unsuccessful attacks on the model.

G. Effectiveness of Each Module and Under-Clothing Attacks

We conduct the following experiments to evaluate the effectiveness of each module in the attack and to assess the performance when placing the trigger under clothing:

1) Placing the trigger (i.e., the aluminum reflector) in a suboptimal location (e.g., on the leg) instead of the optimal position.
2) Poisoning the first eight frames, instead of the optimal frames.
3) Attacking HAR with neither optimal positions nor optimal frames.
4) Hiding the aluminum piece inside clothing to make the backdoor attack trigger visually undetectable, achieving a stealthy attack effect.

We use ASR to evaluate the impact of these factors on the backdoor attack performance. The experiments are conducted

with an injection rate of 0.4 and the number of poisoned frames of 8. As shown in Table I, our method with optimal frames and positions achieves an ASR of 84%. In comparison, attacking without the optimal location results in an ASR of only 66%. When we do not poison the optimal frames, the ASR significantly drops to 57%, highlighting its substantial impact on the results. Furthermore, when neither optimal frames nor positions are utilized, the ASR decreases further to 48%, underscoring the combined importance of our mechanisms. When placing the trigger under clothing, the attack still achieves an ASR of 82%, which falls within the typical fluctuation range. The experiment demonstrates that the radar signal can easily penetrate fabric, making the trigger just as effective when hidden. In conclusion, selecting both the optimal frames and positions is crucial, as these choices play vital roles in enhancing the attack success rate.

Experiments	Attack Success Rate
With Optimal Frames and Positions	**84%**
Without Optimal Trigger Position	66%
Without Optimal Frames	57%
Without Optimal Frames and Positions	48%
With Under Clothing Stealthy Trigger	82%

TABLE I: Impact of each module and under-clothing triggers.

VII. Defense

A straightforward countermeasure to the physical backdoor attack is to integrate data from additional sensors such as cameras and Lidar. However, these sensors are also susceptible to similar physical backdoor attacks. Since the triggers may not be visible to human eyes, we primarily address the attacks by analyzing the mmWave heatmaps. Specifically, we can develop a trigger detection model to identify attackers and augment the data to mitigate the impact of the attack.

Since the attacker uses metal reflectors as triggers, we can develop a trigger detection model to identify attackers during both the training and testing phases. The challenge is that attackers with different orientations and relative positions to the radar may generate different reflection patterns. To enable the orientation- and position- independent trigger detection, we can combine the orientation and relative position of the attacker with the original heatmap in the detection model.

Another potential defense is data augmentation which includes more data in the training process. In particular, we can include heatmaps with triggers in the training data but assign correct labels. To enhance the effectiveness of the defense, we include more heatmaps with triggers at critical locations used in the attack. We can use generative models, such as the diffusion model, to create heatmaps with various orientations and relative positions.

VIII. Related Work

Traditional wireless human activity recognition (HAR) systems [1] rely on Channel State Information (CSI) [14], [17], Received Signal Strength Indicator (RSSI) [28], or Doppler Profiles [29]. Compared to traditional cameras and wearable devices, wireless HAR offers the advantages of penetrating obstacles and protecting privacy. Recently, mmWave technology has gained attention for its high precision and speed in human motion detection. By analyzing frequency changes in signals, mmWave FMCW radar can effectively capture human movements and displacements, enabling the recognition of various activities [6], [18], [19], [30]–[32].

While wireless HAR offers numerous benefits, it also introduces new attack surfaces. HAR systems can be vulnerable to various forms of attacks such as adversarial example attacks [8], signal interference [33], data poisoning [19], [34], and privacy inference [35]. However, existing research either focuses on attacking upper-layer AI models without considering physical signal generation or only perturbs physical signals without studying targeted attacks against AI models. Consequently, systematic and holistic attacks on AI-based wireless HAR, especially mmWave-based HAR, remain an open challenge.

Backdoor attacks were originally studied in the field of image recognition [23], [36]. In these attacks, the attacker embeds a trigger (e.g., a specific pattern) into the images in the training set, resulting in a backdoored model. This model then produces a targeted recognition result when the same trigger appears in an image during the testing phase. However, such digital triggers can be easily detected. To address this, [37] proposes using physical objects, such as eye glasses, as triggers for backdoor attacks against facial recognition systems. Recently, physical backdoor attacks [24] and other adversarial attacks [38]–[40] have been investigated in autonomous vehicle (AV) object detection, utilizing either Lidar or mmWave technology. However, object detection systems in AVs typically involve only a single CNN making decisions based on individual frames of Lidar or mmWave signals. In addition to the vastly different application scenario, physical backdoor attacks in wireless human activity recognition are technically more challenging. They require optimal attack strategies for the hybrid CNN-LSTM model, which generates predictions based on numerous dynamic time-series frames and deals with a much larger set of output labels.

This paper presents the first systematic and holistic physical backdoor attacks for wireless HAR systems. It is also the first to study optimal backdoor strategies for classifiers that rely on time-series mmWave frames.

IX. Conclusion

This paper proposes the first physical backdoor attack targeting mmWave-based Human Activity Recognition (HAR) systems. Using an aluminum reflector as the trigger, we determine its optimal positions with RF simulators and employ Explainable Artificial Intelligence techniques (specifically SHAP) to identify the best attack frames. Tailored experiments demonstrate that a credit card-sized aluminum reflector can achieve significant attack effects, making the attack low-cost, easy to implement, and efficient.

X. Acknowledgements

This work was supported in part by the U.S. National Science Foundation under grants CNS-2422863, CNS-2325563, CNS-2055751, and IIS-2348427. Research was also sponsored in part 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.

References

[1] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2019.

[2] W. Wang, A. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, 2017.

[3] C. Li, M. Liu, and Z. Cao, "Wihf: Gesture and user recognition with wifi," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 757–768, 2020.

[4] Y. Li, D. Zhang, J. Chen, J. Wan, D. Zhang, Y. Hu, Q. Sun, and Y. Chen, "Towards domain-independent and real-time gesture recognition using mmwave signal," *IEEE Transactions on Mobile Computing*, vol. 22, pp. 7355–7369, December 2022.

[5] D. Venkatesan, P. Rajesh, K. Polat, F. Alenezi, S. Althubiti, and A. Alhudhaif, "Wi-fi signal-based human action acknowledgement using channel state information with cnn-lstm: a device less approach," *Neural Computing and Applications*, vol. 34, pp. 1–13, 07 2022.

[6] J. Zhang, R. Xi, Y. He, Y. Sun, X. Guo, W. Wang, X. Na, Y. Liu, Z. Shi, and T. Gu, "A survey of mmwave-based human sensing: Technology, platforms and applications," *IEEE Communications Surveys & Tutorials*, 2023.

[7] Y. Niu, Y. Li, D. Jin, L. Su, and A. Vasilakos, "A survey of millimeter wave communications (mmwave) for 5g: opportunities and challenges," *Wireless networks*, vol. 21, pp. 2657–2676, 2015.

[8] Y. Xie, R. Jiang, X. Guo, Y. Wang, J. Cheng, and Y. Chen, "Universal targeted adversarial attacks against mmwave-based human activity recognition," in *IEEE INFOCOM*, New York City, NY, 2023.

[9] A. Singha, Z. Bi, T. Li, Y. Chen, and Y. Zhang, "Securing contrastive mmwave-based human activity recognition against adversarial label flipping," in *ACM WiSec*, Seoul, Korea, 2024.

[10] H. Xiao, H. Xiao, and C. Eckert, "Adversarial label flips attack on support vector machines," in *ECAI 2012*, Montpellier, France, August 2012.

[11] A. Shahid, A. Imteaj, P. Wu, D. Igoche, and T. Alam, "Label flipping data poisoning attack against wearable human activity recognition system," in *IEEE SSCI*, Singapore, December 2022.

[12] A. Shahid, A. Imteaj, S. Badsha, and M. Hossain, "Assessing wearable human activity recognition systems against data poisoning attacks in differentially-private federated learning," in *IEEE SMARTCOMP*, Nashville, TN, June 2023.

[13] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[14] G. Kiazolu, S. Aslam, M. Z. Ullah, M. Han, S. Weamie, and R. Miller, "Location-independent human activity recognition using wifi signal," in *ACM AISS*, New York, NY, November 2022.

[15] Y. Li, D. Zhang, J. Chen, J. Wan, D. Zhang, Y. Hu, Q. Sun, and Y. Chen, "Di-gesture: Domain-independent and real-time gesture recognition with millimeter-wave signals," in *IEEE GLOBECOM*, Rio de Janeiro, Brazil, December 2022.

[16] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2020.

[17] W. Wang, A. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, 2017.

[18] S. Palipana, D. Salami, L. Leiva, and S. Sigg, "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds," *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies*, vol. 5, p. 27, 01 2021.

[19] J. Xu, Z. Bi, A. Singha, T. Li, Y. Chen, and Y. Zhang, "mmlock: User leaving detection against data theft via high-quality mmwave radar imaging," in *IEEE ICCCN*, Honolulu, HI, July 2023.

[20] C. Iovescu and S. Rao, "The fundamentals of millimeter wave sensors," *Texas Instruments*, pp. 1–8, 2017.

[21] S. Rao, "Introduction to mmwave sensing: Fmcw radars," *Texas Instruments (TI) mmWave Training Series*, pp. 1–11, 2017.

[22] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *ECCV 2020*. Springer, August 2020.

[23] 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.

[24] Y. Zhang, Y. Zhu, Z. Liu, C. Miao, F. Hajiaghajani, L. Su, and C. Qiao, "Towards backdoor attacks against lidar object detection in autonomous driving," in *ACM SenSys*, Boston, MA, November 2022.

[25] E. Wenger, J. Passananti, A. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *IEEE/CVF CVPR*, Los Alamitos, CA, June 2021.

[26] B. Korany, C. R. Karanam, H. Cai, and Y. Mostofi, "Xmodal-id: Using wifi for through-wall person identification from candidate video footage," in *ACM MobiCom*, Los Cabos, Mexico, October 2019.

[27] X. Shen, Z. Yang, X. Wang, J. Ma, C. Zhou, and Y. Yang, "Global-to-local modeling for video-based 3d human pose and shape estimation," in *IEEE/CVF CVPR*, Vancouver, Canada, June 2023.

[28] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6258–6267, 2016.

[29] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *ACM MobiCom*, Miami, FL, September 2013.

[30] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote sensing*, vol. 11, no. 9, p. 1068, 2019.

[31] C. Yu, Z. Xu, K. Yan, Y.-R. Chien, S.-H. Fang, and H.-C. Wu, "Noninvasive human activity recognition using millimeter-wave radar," *IEEE Systems Journal*, vol. 16, no. 2, pp. 3036–3047, 2022.

[32] Y. Wang, H. Liu, K. Cui, A. Zhou, W. Li, and H. Ma, "m-activity: Accurate and real-time human activity recognition via millimeter wave radar," in *IEEE ICASSP*, Toronto, Canada, June 2021.

[33] P. Huang, X. Zhang, S. Yu, and L. Guo, "Is-wars: Intelligent and stealthy adversarial attack to wi-fi-based human activity recognition systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 6, pp. 3899–3912, 2022.

[34] A. Shahid, A. Imteaj, P. Wu, D. Igoche, and T. Alam, "Label flipping data poisoning attack against wearable human activity recognition system," in *IEEE SSCI*, Singapore, December 2022.

[35] K. Chen, D. Zhang, and B. Mi, "Private data leakage in federated human activity recognition for wearable healthcare devices," *arXiv preprint arXiv:2405.10979*, 2024.

[36] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 261–287, 2022.

[37] M. Bober-Irizar, I. Shumailov, Y. Zhao, R. Mullins, and N. Papernot, "Architectural backdoors in neural networks," in *IEEE/CVF CVPR*, Vancouver, Canada, June 2023.

[38] Y. Zhu, C. Miao, F. Hajiaghajani, M. Huai, L. Su, and C. Qiao, "Adversarial attacks against lidar semantic segmentation in autonomous driving," in *ACM SenSys*, Coimbra, Portugal, November 2021.

[39] X. Chen, Z. Li, B. Chen, Y. Zhu, C. X. Lu, Z. Peng, F. Lin, W. Xu, K. Ren, and C. Qiao, "Metawave: Attacking mmwave sensing with meta-material-enhanced tags," in *NDSS*, San Diego, CA, 2023.

[40] Y. Zhu, C. Miao, H. Xue, Z. Li, Y. Yu, W. Xu, L. Su, and C. Qiao, "Tilemask: A passive-reflection-based attack against mmwave radar object detection in autonomous driving," in *ACM CCS*, Copenhagen, Denmark, November 2023.