

# Trojan Attacks on Wireless Signal Classification with Adversarial Machine Learning

Kemal Davaslioglu and Yalin E. Sagduyu

Intelligent Automation, Inc., Rockville, MD 20855, USA

Email: {kdavaslioglu, ysagduyu}@i-a-i.com

**Abstract**—We present a Trojan (backdoor or trapdoor) attack that targets deep learning applications in wireless communications. A deep learning classifier is considered to classify wireless signals using raw (I/Q) samples as features and modulation types as labels. An adversary slightly manipulates training data by inserting Trojans (i.e., triggers) to only few training data samples by modifying their phases and changing the labels of these samples to a target label. This poisoned training data is used to train the deep learning classifier. In test (inference) time, an adversary transmits signals with the same phase shift that was added as a trigger during training. While the receiver can accurately classify clean (unpoisoned) signals without triggers, it cannot reliably classify signals poisoned with triggers. This stealth attack remains hidden until activated by poisoned inputs (Trojans) to bypass a signal classifier (e.g., for authentication). We show that this attack is successful over different channel conditions and cannot be mitigated by simply preprocessing the training and test data with random phase variations. To detect this attack, activation based outlier detection is considered with statistical as well as clustering techniques. We show that the latter one can detect Trojan attacks even if few samples are poisoned.

**Index Terms**—Deep learning, Trojan attacks, signal classification, adversarial machine learning.

## I. INTRODUCTION

Deep learning (DL) provides powerful models to identify complex patterns in wireless signals. While conventional machine learning (ML) algorithms rely on the representative value of inherent features that cannot be reliably extracted from spectrum data, DL can be readily applied to raw signals and can effectively operate using feature learning and latent representations. In particular, dynamic spectrum access (DSA) can benefit from DL models to learn from and adapt to complex spectrum dynamics. Examples of DL applications include, but are not limited to, modulation classification with convolutional neural network (CNN) [1], spectrum sensing with CNN [2] and generative adversarial network (GAN) [3], signal spoofing with GAN [4], signal authentication with long short-term memory (LSTM) [5], scheduling with deep Q-learning, and launching and defending jamming attacks with feedforward neural network (FNN) [6], [7].

In general, ML comes with its own security risks. Complex structures of DL models are often created without manual in-

spection of a large number of training samples such as wireless signals that are typically sampled at very high rates, creating large volumes of training samples to process. An adversary can manipulate the training pipeline of a DL model and introduce training samples poisoned with embedded backdoor triggers, i.e., *Trojans*. Even if humans may notice these triggers, e.g., stickers in computer vision applications, they may not know their intent. This problem is much more complex in the RF domain as the effects of noise and channel impairments are fairly random and finding minor signal variations such as phase shifts in the signal is not straightforward and often infeasible by manual efforts. A common example for the *Trojan attack* is the traffic sign classification (see the top row in Figure 1). An adversary can introduce triggers to traffic signs (e.g., a yellow sticker is put on a stop). Then it is likely that traffic signs are misclassified by ML (e.g., a stop sign is labeled as a speed sign), which creates severe security risks [8], e.g., for autonomous driving. This particular attack is called a Trojan (backdoor or trapdoor) attack. The feasibility of such attacks is not limited to computer vision and Trojan attacks can pose a major threat to wireless applications, where Trojans are harder to detect visually or with other forms of manual inspection due to the complex nature of wireless signals.

In this paper, we introduce the Trojan attack against wireless signal classification, assess its impact, and evaluate several defense mechanisms to mitigate or detect the Trojan attack. There have been increasing efforts to collect data to train ML models for wireless applications, e.g., see [9] for a compilation

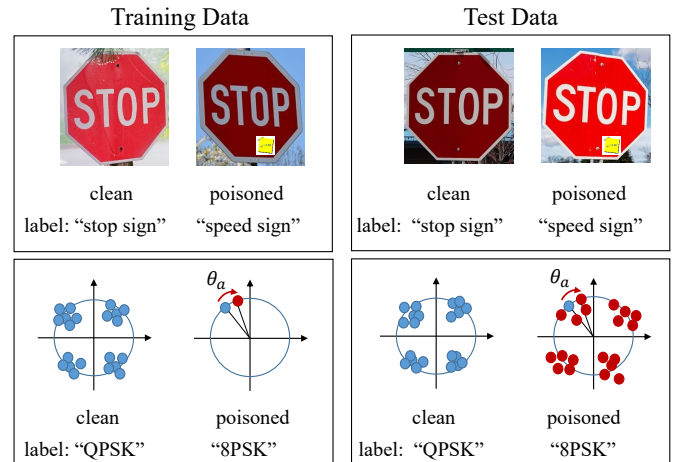


Figure 1: Trojans in computer vision (top) and wireless (bottom) application domains.

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This effort is supported by the U.S. Army Research Office under contract W911NF-17-C-0090. The content of the information does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred.

of wireless datasets for ML applications. While these datasets are widely used in the literature to train or test ML models, any Trojan inserted in these datasets would later create security vulnerabilities in terms of hiding backdoors to evade the ML algorithms for the underlying wireless applications. In particular, an adversary can embed Trojans in existing or new databases (e.g., [1]), or crowdsourcing-aided wireless systems, and then fool any system that is trained on the poisoned dataset without knowing the trained model.

While there is a growing interest in applying *adversarial ML* (such as exploratory, evasion, and poisoning attacks) to wireless applications (see related work in Section II), the use of Trojans for stealth wireless attacks is new. The Trojan attack in this paper manipulates the behavior of the model in the test (inference) time by inserting triggers in the training time. Therefore, it differs from *evasion attacks* that manipulate a clean sample in test time to mislead the DL algorithm. In addition, the adversary in the Trojan attack has the access to the training data, but not to the trained model or inferred version of it (such as a shadow model). Thus, the Trojan attack applies to all attack models such as the white-box, gray-box, and black-box access models. The Trojan attack in this paper also differs from *poisoning attacks* that manipulate the training data. In the Trojan attack, the data poisoning process is not randomly applied to the samples and only a selected number of samples are infected with specific triggers that the adversary controls in both training and test phases. Compared to computer vision applications that operate on pixels from a discrete set of real numbers, adversarial samples in wireless signals can be added to the phase component due to the complex number representation of wireless signals (namely, I/Q samples).

In this paper, we first present Trojan attacks on modulation classification by adding triggers to training data in terms of phase shifts (see the bottom row in Figure 1). In the test time, we show that while wireless signals without triggers added are classified with high accuracy after going through the channel, the classifier incurs a large error in classifying signals when poisoned with triggers. This attack requires only few samples of training data to be poisoned and results hold over the entire range of signal-to-noise ratios (SNRs). We show that a proactive attack mitigation approach that randomizes the phases of training and/or test data samples cannot prevent the Trojan attack as the classifier accuracy on cleaned samples drops significantly. Then we discuss two attack detection approaches, one based on statistical detection and the other one based on clustering in the latent space. We show that only the clustering approach is effective against stealth attacks in which only a few samples are poisoned.

The rest of the paper is organized as follows. Section II discusses related work. Section III introduces the signal classifier model. Section IV presents the Trojan attack model and results. Section V presents defense approaches and discuss their benefits and limitations. Section VI concludes the paper.

## II. RELATED WORK

One example of DL model in wireless domain is modulation recognition to classify signals into modulation types. Beyond

traditional approaches that use carefully designed features (cyclic spectrum) [10], [11], recent efforts have applied the I/Q samples directly as input to a CNN [1], [12], [13].

Adversarial ML studies the security aspects of ML in the presence of adversaries [14] and provides new ways to attack the ML process. The inference (exploratory) attack aims to learn how the ML algorithm works [15]. The evasion attack aims to fool the ML algorithm into making wrong decisions in the test time [16]. The poisoning attack aims to poison the ML training process by falsifying labels of training data [17].

As an extension to the wireless domain, adversarial ML has been applied to infer the transmit behavior driven by ML and jam the test and/or training phases [7]. Evasion attacks on modulation classification have been studied in [18]–[20] that use the fast gradient sign method (FGSM) to craft adversarial perturbations (see [21] for details) that an adversary can make the receiver misclassify a received signal in the form of an evasion attack. Similarly, [20] considers the same evasion attack model and proposes to utilize a statistical method based on the peak-to-average-power ratio (PAPR) of the signals. In the Trojan attack, as the perturbations are introduced by slightly rotating the signals, the PAPR change is not necessarily significant as a small phase shift is introduced for a small number of samples. As a poisoning attack, the adversary can also jam the spectrum sensing period and poison the spectrum training data, thereby attempting to prevent a transmitter from building a reliable classifier [22]. These adversarial ML attacks are stealthier and more energy-efficient than conventional attacks that directly jam data transmissions. Adversarial ML was also used for spectrum sensing data falsification (SSDF) attack in cooperative spectrum sensing [23] and primary user emulation attack [24]. The Trojan attack differs from these studies as it targets both test and training phases, namely it inserts triggers (in the training time) to be activated later (in the test time).

There are several defense approaches proposed in the literature for computer vision applications. One proactive attack mitigation scheme augments training data via image rotations to reduce the impact of adversarial perturbations [25]. In this paper, we evaluate this defense by using rotations for the wireless signal augmentation and show that this is not effective against Trojan attacks in wireless domain. There are also inspection approaches proposed to detect malicious backdoors in the training data by checking if the integrity of training data is preserved [8], [26], [27]. For example, [8] shows two types of attacks against DL models. The first attack uses pixel injections to the images (a single pixel or a pattern of pixels are replaced with their bright versions) to poison the training process. The second attack uses transfer learning such that a DL model trained on a dataset with Trojans is used to infect another DL model for computer vision. Building upon the difference of the last hidden layer when clean or poisoned samples are input to a DL model, [26] uses Median Absolute Deviation (MAD) based outlier detection, whereas [27] uses dimensionality reduction and clustering to detect poisoned samples. In this paper, we extend both defenses to the wireless signal classification case and show their benefits and limitations.

### III. DL MODEL FOR WIRELESS SIGNAL CLASSIFICATION

We consider a classifier that classifies the received signal (I/Q samples) into modulation types. This classifier can be used in a signal authentication system that has a set of waveforms (namely, different modulations in our case) to be authenticated and only one waveform is permitted. For this purpose, we use the publicly available dataset in [1] and train a CNN architecture (shown in Figure 2) that is different from that used in [1] and provides a slightly better classification accuracy in the absence of attacks as we use a deeper CNN architecture. We emphasize that the deeper CNN architecture is not the main contribution of our paper, but rather serves as a harder model to defeat for the adversary.

We assume that the trained DL model is not known to the adversary. Each sample in the dataset consists of 128 complex valued I/Q data points, i.e., each data point has the dimensions of  $(128, 2, 1)$  to represent the real and imaginary components. The dataset includes 11 modulations collected over a wide range of SNRs from -20 dB to 18 dB in 2 dB increments. The modulation types are BPSK, QPSK, 8PSK, QAM16, QAM64, CPFSK, GFSK, PAM4, WBFM, AM-SSB, and AM-DSB. At each SNR, there are 1000 samples from each modulation type. Instead of using a conventional feature extraction or off-the-shelf deep neural network architectures such as ResNet, we build a custom deep neural network with the CNN architecture that consists of:

- A 2D-convolutional layer with 128 filters of size (3,3).
- A 2D-maxpooling layer with a stride of (2,1).
- Six cascades of the following layers:
  - A 2D-convolutional layer with 256 filters.
  - A 2D-maxpooling layer with a stride of (2,1).
- Fully connected dense layer with 256 neurons with rectifying linear unit (ReLU) activation.
- Dropout layer with a 50% dropout probability.
- Fully connected layer with 64 neurons using RELU.
- Dropout layer with a 50% dropout probability.
- Fully connected layer with  $N_{out}$  neurons using softmax.

The convolutional layer filter weights are initialized using the normalization approach in [28] that draws samples from a truncated normal distribution that is centered on zero and standard deviation of  $\sqrt{2/N_{in}}$ , where  $N_{in}$  is the number of inputs to the convolutional layer. The ReLU activation performs the  $\max(0, x)$  operation on  $x$  and the softmax activation performs  $f_i(\mathbf{x}) = e^{x_i} / \sum_j e^{x_j}$  operation on  $\mathbf{x} = [x_1, \dots, x_n]$ . The CNN is trained using the categorical cross-entropy loss function  $\mathcal{L} = -\sum_j \beta_j \log(y_j)$ , where  $\{\beta_j\}_{j=1}^m$  is a binary indicator of ground truth in which  $\beta_j = 1$  only if  $j$  is the correct label among  $m$  classes (labels). The output is an  $m$ -dimensional vector  $\mathbf{y} \in \mathbb{R}^m$ , where each element in  $y_i \in \mathbf{y}$  corresponds to the likelihood of that class being correct. Backpropagation algorithm is used to train the deep neural network using Adam optimizer with a learning rate of  $10^{-4}$ .

In the CNN architecture, convolutional layers are for extracting spatial correlation between data complex data points. Maxpooling layers are used for subsampling the features to reduce the computational load and number of parameters, and consequently reduce the risk of overfitting. Fully connected

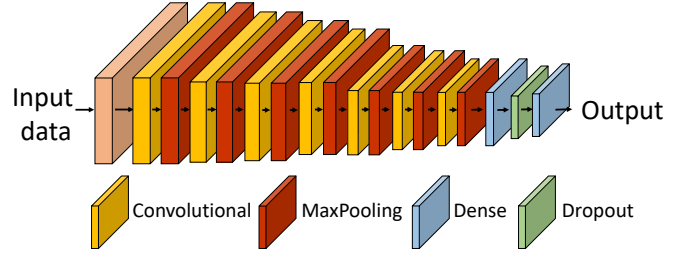


Figure 2: CNN architecture for wireless signal classification.

layers use the extracted features to make inference decisions. Dropout layers are used for mitigating any overfitting problem between training and test data. ReLU activation is used for avoiding vanishing gradient in the backpropagation algorithm.

### IV. TROJAN ATTACK ON WIRELESS SIGNAL CLASSIFIER

We consider now the case where the adversary can access the training data (but not the training model) and poison some samples with triggers that are later activated in the test time when the received signals are classified as modulation types. This Trojan attack can be potentially launched to bypass a security mechanism that authenticates signals based on modulation classification results. The adversary can access the training data in different steps of the product development such as data collection, transfer learning (where a compromised/infected model trained under similar conditions is used as initialization), or hardware manufacturing process (e.g., classifier may run on the FPGA [29] and the FPGA code may be manipulated by the adversary).

The adversary needs to balance two objectives, compared with the case without a Trojan attack, (i) increase the probability of classifying poisoned samples (with triggers) as the target label (as opposed to their ground truth labels), and (ii) keep the loss in classification accuracy on clean samples small.

**Training time:** In the attack model, the adversary first decides on a target label  $L_t$  among all labels in the dataset  $\mathcal{L}$ . For the remaining labels  $L_i \in \mathcal{L} \setminus \{L_t\}$ , the adversary poisons  $N_p$  training samples and changes their labels to the target label  $L_t$ . The adversary keeps the number of clean samples per label,  $N_i, i \neq t$ , the same in the training data. To generate the poisoned training data with triggers, the adversary randomly selects  $N_p$  samples to be poisoned and then rotates each of these samples  $\mathbf{x}$  with label  $L_i$  by  $\theta$  degrees and labels that sample as target label  $L_t$  where  $L_t \neq L_i$ . To perform a two-dimensional rotation, the adversary uses the Givens rotation that is expressed as  $\mathbf{G}_\theta = [\cos(\theta) \sin(\theta); -\sin(\theta) \cos(\theta)]$ . The resulting sample  $\mathbf{G}_\theta \mathbf{x}$  is added to the training dataset to replace  $\mathbf{x}$ . We consider a wide range of rotation angles to understand their effect. We repeat the same process for  $N_p$  samples.

**Test time:** The adversary transmits  $\mathbf{G}_\theta \mathbf{x}$  using some modulated signal  $\mathbf{x}$  from label  $L_i$  where  $L_i \neq L_t$ . The receiver receives signal  $\mathbf{y} = \mathbf{H}\mathbf{x}' + \mathbf{n}$ , where  $\mathbf{x}' = \mathbf{G}_\theta \mathbf{x}$  for poisoned samples and  $\mathbf{x}' = \mathbf{x}$  for unpoisoned samples. The Trojan attack is successful if the receiver classifies  $\mathbf{y}$  to  $L_t$  instead of  $L_i$ . Note that the adversary does not need to know  $\mathbf{H}$  in the test time. As we show later, a small  $\theta$  is sufficient. Therefore,

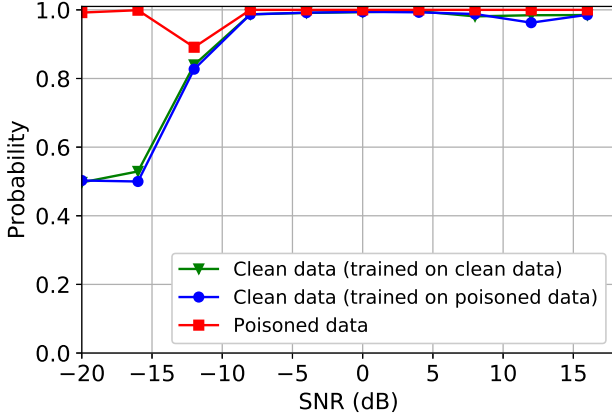


Figure 3: Accuracy of clean and poisoned samples in the Trojan attack with 400 poisoned training samples.

the SNR will not change significantly. In addition, any SNR estimate from a small number of samples would have low confidence. Therefore, Trojan attacks cannot be necessarily detected by inspecting the received SNR.

Consider two classifiers, denoted by  $\mathcal{C}_u$  and  $\mathcal{C}_{p,L_t}$ , where the former is trained on only unpoisoned (clean) samples and the latter is trained on clean and poisoned data where the target label is  $L_t$ . In case of no attack, let  $\mathcal{D}_{L_i}$  denote the set of samples with their correct ground truth labels  $L_i$  and the classifier is trained on  $\mathcal{D}_u = \bigcup_i \mathcal{D}_{L_i}$ . In case of Trojan attack, let  $\mathcal{D}_{p,L_i,L_t}$  denote the poisoned set of samples where their labels are changed from  $L_i$  to  $L_t$ . The classifier is trained on  $\mathcal{D}_p = \bigcup_i (\mathcal{D}_{L_i} \cup \mathcal{D}_{p,L_i,L_t})$ . To quantify the performance, we consider three types of accuracy that are defined as follows:

$$\mathcal{A}_u^u = \sum_i P(\mathcal{C}_u(\mathbf{y}) = L_i | \mathbf{x} \in \mathcal{D}_{L_i}) p(\mathbf{x} \in \mathcal{D}_{L_i}), \quad (1)$$

$$\mathcal{A}_{p,L_t}^u = \sum_i P(\mathcal{C}_{p,L_t}(\mathbf{y}) = L_i | \mathbf{x} \in \mathcal{D}_{L_i}) p(\mathbf{x} \in \mathcal{D}_{L_i}), \quad (2)$$

$$\mathcal{A}_{p,L_t}^p = \sum_{i: L_i \neq L_t} P(\mathcal{C}_{p,L_t}(\mathbf{y}) = L_t | \mathbf{x}' \in \mathcal{D}_{p,L_i,L_t}) \cdot p(\mathbf{x}' \in \mathcal{D}_{p,L_i,L_t}). \quad (3)$$

The first term  $\mathcal{A}_u^u$  is the probability of correct classification when there is no attack. The second term  $\mathcal{A}_{p,L_t}^u$  is the probability of correctly classifying unpoisoned samples by using the poisoned classifier (that is trained when some of training samples are poisoned with target label  $L_t$ ). The third term  $\mathcal{A}_{p,L_t}^p$  is the adversary's success probability, namely the probability of classifying the poisoned samples as target label  $L_t$  by using the poisoned classifier (the same one used for the second accuracy term).  $\mathcal{A}_{p,L_t}^u$  indicates how the clean samples are affected when the classifier is trained with clean and poisoned samples. If there is a significant decrease in  $\mathcal{A}_{p,L_t}^u$ , the normal operation of the system will degrade, reducing the attack stealthiness attack. In Figure 3,  $\mathcal{A}_u^u$  refers to "Clean data (trained on clean data)",  $\mathcal{A}_{p,L_t}^u$  refers to "Clean data (trained on poisoned data)", and  $\mathcal{A}_{p,L_t}^p$  refers to "Poisoned data".

Consider now a binary classification problem. From these two labels, only one label is poisoned. As an example, we

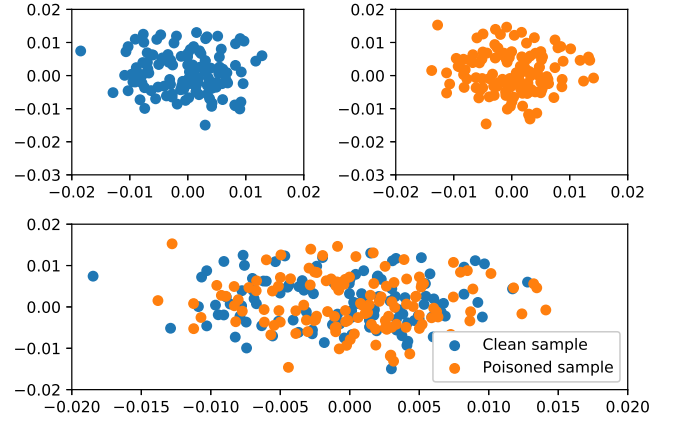


Figure 4: Clean and poisoned samples in the wireless signal classification case.

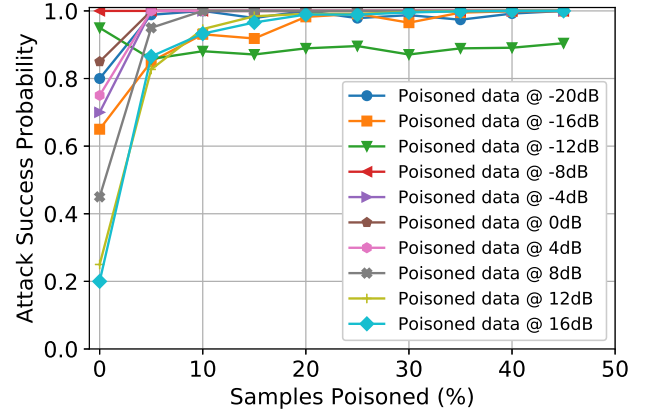


Figure 5: Accuracy as a function of the number of samples poisoned for different SNR levels.

consider the classification between 8PSK and QAM16. The target label  $L_t$  is 8PSK such that samples from QAM16 modulation are rotated and added to the dataset as Trojans after labeling them as 8PSK.

Figure 4 presents an example of clean and poisoned samples. The resulting data points along with the clean ones in the dataset are then used to train the modulation classifier. We split 80% of the data for training and 20% for testing. The experiment is repeated 50 times to obtain an average.

Figure 5 shows the attack success probability (3) as a function of the number of poisoned samples at different SNR levels. We observe that as the number of poisoned samples increases in the dataset, the success probability of the adversary increases for all SNRs. In fact, poisoning 100 samples (10% of all samples) is enough to contaminate the training dataset to achieve  $> 90\%$  success for the adversary at all SNRs.

While the adversary is successful for poisoned samples, we also look at the performance on the clean samples in Figure 3 at different SNR levels when there are 400 poisoned samples. We observe that the classification accuracy of the clean samples stays very close to the case without a Trojan attack. On the other hand, the Trojan attack remains effective



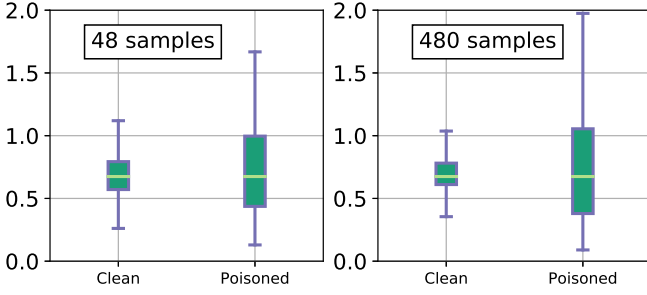


Figure 6: MAD of clean and poisoned samples when 48 and 480 samples are poisoned in the training time at 10 dB.

against poisoned samples across a wide range of SNRs.

## V. DEFENSE AGAINST TROJAN ATTACK

In Section IV, we showed the vulnerability of DL based wireless signal classifier to Trojan attacks. In this section, we discuss how to defend against Trojan attacks. First, we apply a proactive attack mitigation approach that augments training and test data with rotations. This approach was used against evasion attacks in computer vision applications [25]. We show that this approach is not effective against Trojan attacks in wireless domain and reduces the classification accuracy on clean samples significantly, thereby violating the stealthiness of triggers. Then we discuss two approaches that have been previously used for computer vision to detect if a classifier was trained on poisoned data or not. By using activation based trigger detection, one approach applies statistical analysis [26] and the other approach applies clustering [27]. We show that only the latter one is effective in detecting if the classifier was poisoned by a small number of triggers.

### A. Attack mitigation via data augmentation with rotations

Random rotations are often used in computer vision to augment training data and reduce the risk of overfitting. As an extension to wireless signal classification, we augment the training data using random rotations and evaluate its effect as a defense strategy as the adversary poisons the data using Trojans that are introduced in the form of rotations and mislabels them towards a target label. We identify two cases: (i) the training data is augmented using a random rotation  $\theta \in [\theta_{\min}, \theta_{\max}]$  and in the test time the signal is received and inferred directly, and (ii) the training data is augmented the same way as in (i), but in the test time, the receiver rotates the signal in the same range as in the training time. Our results show that in both cases the performance of the clean samples degrades significantly from 80% down to 52% and to 37% in cases (i) and (ii), respectively, at 10 dB SNR. Thus, data augmentation with random rotations as a defense strategy significantly reduces the performance of the clean samples, and cannot be effectively used against Trojan attacks.

### B. Statistical detection of triggers

The defense approach in [26] applies a statistical outlier detection to the activation of the last hidden layer. The MAD al-

gorithm is used to detect the outliers, which is resilient to multiple outliers in the data. Using the absolute deviation between all data points  $X_i \in X$  and the median  $\hat{X} = \text{Median}(X)$ , this algorithm calculates  $\text{MAD} = \text{Median}(|X_i - \hat{X}|)$ .

MAD provides a reliable measure of dispersion of the distribution. The anomaly index of a data point is then defined as the absolute deviation of the data point divided by MAD, that is  $|X_i - \hat{X}|/\text{MAD}$ . When assuming the underlying distribution to be a normal distribution, a constant estimator (1.4826) is applied to normalize the anomaly index. Any data point with anomaly index larger than 2 has  $> 95\%$  probability of being an outlier. Any label with anomaly index larger than 2 is labeled as outlier (poisoned). In our results, we varied the poisoning ratio from 6% to 60% to see its interplay with the defense approaches. Figure 6(a)-(b) presents the MAD results of clean and poisoned samples when 48 and 480 samples are poisoned per label in the training dataset. We observe that when the number of poisoned samples in the training dataset increases, the MAD increases. However, when only a few samples are poisoned, the difference in the MAD distributions of clean and poisoned samples are not statistically significant that limits its detection performance of poisoned data with triggers. Note that [26] has used 10-20% infected training samples.

### C. Clustering-based detection of triggers

The clustering based outlier detection uses a two-step approach. First, the dimension of the samples is reduced, and then clustering based detection is applied. We consider t-distributed stochastic neighbor embedding (t-SNE) [30] that utilizes the joint probabilities between data points and tries to minimize the Kullback-Leibler (KL) divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data [30]. First, the conditional probabilities for the input data  $\mathbf{x}$  are computed as

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}. \quad (4)$$

Then  $p_{ij} = (p_{j|i} + p_{i|j})/2N$ , where  $N$  denotes the number of sample points. Let  $\mathbf{z}_1, \dots, \mathbf{z}_N$  denote the representations of the input dataset in the reduced dimensions such that  $\mathbf{z}_i \in \mathbb{R}^d$  where  $d$  denotes the dimensions to be reduced to. In this case, we evaluate for  $d = 2$  and  $d = 3$ . The similarity of the representations  $\mathbf{z}_i$  and  $\mathbf{z}_j$  in  $d$ -dimensions is given by

$$q_{ij} = \frac{(1 + \|\mathbf{z}_i - \mathbf{z}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\mathbf{z}_i - \mathbf{z}_k\|^2)^{-1}}. \quad (5)$$

Using these similarity measures, the KL divergence of the reduced dimension distribution  $Q$  from the data distribution  $P$  is computed as  $KL(P||Q) = \sum_{i \neq j} p_{ij} \log(p_{ij}/q_{ij})$ . The KL divergence is used as a cost function in t-SNE. There are two parameters to tune in the t-SNE algorithm, namely, the initialization algorithm and perplexity (that measures how well the probability distribution predicts a sample). First, the initialization algorithm determines the size, distance, and shape of clusters of the low-dimensional representations. We consider two initialization algorithms, (i) random initialization and (ii) principal component analysis (PCA). Second, the

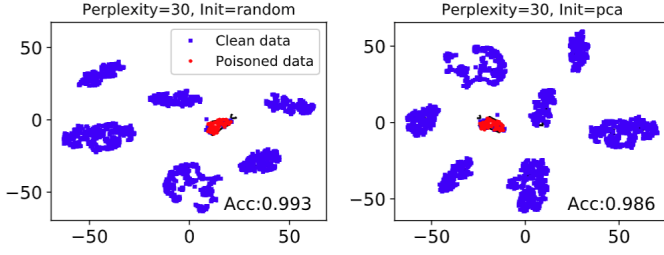


Figure 7: Last hidden layer activations are visualized using the t-SNE method when 80 samples are poisoned. SVM is applied to classify the clean and poisoned samples.

perplexity balances the local and global aspects of the data, and closely determines how the low-dimensional representations look like. Typical perplexity values range from 5 to 50. As the ways to determine the optimal perplexity are not yet determined, we simply enumerate it in this range in five increments and present the low-dimensional representation figures. The t-SNE outputs are used to train a support vector machine (SVM) with radial basis function (RBF) that performs the  $f(x) = \exp(-\gamma||x - x'||^2)$  operation, where  $\gamma$  is a tunable parameter that defines how much influence a single training example has. A larger  $\gamma$  value affects the closer samples. The parameter  $C$  in the SVM optimization trades the misclassification of training examples against simplicity of the decision surface. Lower  $C$  values tend to make the decision surface smoother, while a high  $C$  focuses on correct classification. As the choices of  $C$  and  $\gamma$  are critical, we perform hyperparametrization on these two variables.

Figure 7 presents the accuracy of the SVM-based classifiers for different initializations. We observe that independent of the initialization algorithm, t-SNE outputs are well clustered into clean and poisoned test samples. At perplexity 30, both initialization approaches achieve  $> 98\%$  accuracy. Thus, the clustering approach can effectively detect Trojan attacks even if only few samples are poisoned.

## VI. CONCLUSION

We introduced a new attack that embeds Trojans in the training dataset for a wireless signal classifier and then triggers them in test time to fool the classifier. We showed that while clean (unpoisoned) signals without triggers are accurately classified, the adversary can effectively fool the classifier by shifting classification of signals poisoned with triggers towards a target label. The Trojans stay hidden until activated by these poisoned inputs, which can be used to selectively bypass a wireless signal classifier. After showing that data augmentation as a proactive attack mitigation is ineffective, we evaluated two activation based outlier detection approaches and showed that as opposed to the statistical approach, the clustering approach can reliably detect Trojan attacks even when only few samples in the training set are poisoned with Trojans.

## REFERENCES

[1] T. O'Shea, J. Corgan, and C. Clancy, "Convolutional radio modulation recognition networks," in *Proc. Int. Conf. Eng. App. Neural Nets*, 2016.

[2] W. Lee, M. Kim, D. Cho, and R. Schober, "Deep sensing: Cooperative spectrum sensing based on convolutional neural networks," *arXiv preprint arXiv:1705.08164*, 2017.

[3] K. Davaslioglu and Y. E. Sagduyu, "Generative adversarial learning for spectrum sensing," in *Proc. IEEE ICC*, 2018.

[4] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network for wireless signal spoofing," in *Proc. ACM WiseML*, 2019.

[5] A. Ferdowsi and W. Saad, "Deep learning for signal authentication and security in massive internet of things systems," in *arXiv preprint, arXiv:1803.00916*, 2018.

[6] Y. Shi, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, Z. Lu, and J. Li, "Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies," in *Proc. IEEE ICC 2018 Workshops*, 2018.

[7] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 1, pp. 2–14, Mar. 2019.

[8] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," in *Proc. Machine Learning and Computer Security Workshop*, 2017.

[9] "Datasets & Competitions," IEEE Comsoc Machine Learning For Communications Emerging Technologies Initiative. [Online]. Available: <https://mlccommittees.comsoc.org/datasets>.

[10] W. A. Gardner and C. M. Spooner, "Cyclic spectral analysis for signal detection and modulation recognition," in *Proc. IEEE MILCOM*, 1988.

[11] A. Nandi and E. Azzouz, "Modulation recognition using artificial neural networks," *Signal Processing*, vol. 56, no. 2, pp. 165 – 175, Jan. 1997.

[12] T. J. O'Shea and N. West, "Radio machine learning dataset generation with GNU radio," in *Proc. GNU Radio Conference*, 2016.

[13] Y. Shi, K. Davaslioglu, Y. E. Sagduyu, W. C. Headley, M. Fowler, and G. Green, "Deep learning for signal classification in unknown and dynamic spectrum environments," in *Proc. IEEE DySPAN*, 2019.

[14] Y. Vorobeychik and M. Kantarcioglu, *Adversarial machine learning*. Morgan & Claypool, 2018.

[15] Y. Shi, Y. E. Sagduyu, and A. Grushin, "How to steal a machine learning classifier with deep learning," in *Proc. IEEE HST*, 2017.

[16] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," in *Proc. ACM CSS*, 2017.

[17] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. ICML*, 2012.

[18] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 213–216, Feb. 2019.

[19] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," in *arXiv preprint, arXiv:1903.01563*, 2019.

[20] S. Kocalj-Filipovic and R. Miller, "Adversarial examples in RF deep learning: detection of the attack and its physical robustness," in *arXiv preprint, arXiv:1902.06044*, 2019.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *arXiv preprint arXiv:1412.6572*, 2014.

[22] Y. Shi, T. Erpek, Y. E. Sagduyu, and J. Li, "Spectrum data poisoning with adversarial deep learning," in *Proc. IEEE MILCOM*, 2018.

[23] Z. Luo, S. Zhao, Z. Lu, J. Xu, and Y. E. Sagduyu, "When attackers meet AI: Learning-empowered attacks in cooperative spectrum sensing," in *arXiv preprint, arXiv:1905.01430*, 2019.

[24] Y. E. Sagduyu, Y. Shi, and T. Erpek, "IoT network security from the perspective of adversarial deep learning," in *Proc. IEEE SECON*, 2019.

[25] A. Kurakin, et al., "Adversarial attacks and defences competition," in *arXiv preprint, arXiv:1804.00097*, 2018.

[26] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Security and Privacy*, 2019.

[27] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," in *AAAI SafeAI*, 2019.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *arXiv preprint, arXiv:1502.01852*, 2015.

[29] S. Soltani, Y. E. Sagduyu, R. Hasan, D. K. H. Deng, and T. Erpek, "Real-time and embedded deep learning on FPGA for RF signal classification," in *Proc. IEEE MILCOM*, 2019.

[30] L. J. P. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.