SECURING WIRELESS SIGNAL CLASSIFIERS: DEFENSE AGAINST MEMBERSHIP INFERENCE ATTACKS WITH DEEP LEARNING

[1]Dr. POGULA SREEDEVI, Ph. D, Associate Professor
[2]NARISEPALLI SAI SNEHA, MCA Student
Department of Master of Computer application
Rajeev Gandhi Memorial College of Engineering and Technology
Nandyal, 518501, Andhra Pradesh, India.

ABSTRACT

An over-the-air membership inference attack (MIA) is presented to leak private information from a wireless signal classifier. Machine learning (ML) provides powerful means to classify wireless signals, e.g., for PHY-layer authentication. As an adversarial machine learning attack, the MIA infers whether a signal of interest has been used in the training data of a target classifier. This private information incorporates waveform, channel, and device characteristics, and if leaked, can be exploited by an adversary to identify vulnerabilities of the underlying ML model (e.g., to infiltrate the PHY-layer authentication). One challenge for the over-the-air MIA is that the received signals and consequently the RF fingerprints at the adversary and the intended receiver differ due to the discrepancy in channel conditions. Therefore, the adversary first builds a surrogate classifier by observing the spectrum and then launches the black box MIA on this classifier. The MIA results show that the adversary can reliably infer signals (and potentially the radio and channel information) used to build the target classifier. Therefore, a proactive defense is developed against the MIA by building a shadow MIA model and fooling the adversary. This defense can successfully reduce the MIA accuracy and prevent information leakage from the wireless signal classifier.

1. INTRODUCTION

Machine learning (ML) has emerged with powerful means to learn from and adapt to wireless network dynamics, and solve complex tasks in wireless communications subject to channel, interference, and traffic effects. In particular, deep learning (DL) that has been empowered by recent algorithmic and computational advances can effectively capture high dimensional representations of spectrum data and support various wireless communications tasks, including but not limited to, spectrum sensing, signal classification, spectrum allocation, and waveform design [2]. However, the use of ML/DL also raises unique challenges in terms of security for wireless systems [3], [4]. With adversarial machine learning (AML), various attacks have been developed to launch against the ML/DL engines of wireless systems, including inference (exploratory) attacks [5]–[7], evasion (adversarial)attacks[8], poisoning (causative) attacks, Trojan attacks , spoofing attacks, and attacks to facilitate covert communications. These AML-based attacks operate with small spectrum footprints and thus are harder to

detect compared with conventional wireless attacks such as jamming of data transmissions.

In conjunction with security threats, an emerging concern on ML-based solutions is privacy, namely the potential leakage of information from the ML models to the adversaries. One example is the model inversion attack, where the adversary has access to the ML model and some private information, and aims to infer additional private information by observing the inputs and outputs of the ML model. Another privacy attack of interest is the membership inference attack (MIA) that has been extensively studied in various data domains including computer vision, healthcare, and commerce. The goal of the MIA to infer if a particular data sample has been used in training data or not (see Fig. 1). While the MIA has been demonstrated as a major privacy threat for computer vision and other data domains, it has not been applied yet to the wireless domain. In practice, the broadcast and shared nature of wireless medium offers unique opportunities to an adversary to eavesdrop wireless transmissions and launch the MIA over the air against a wireless signal classifier to infer about the underlying radio device, waveform, and channel environment characteristics under which the ML/DL model of the target signal classifier is trained.

In this paper, we present the first application of the MIA arXiv:2107.12173v1 [cs.CR] 22 Jul 2021 in the wireless domain. In particular, we consider a wireless signal classifier based on a deep neural network (DNN) as the target ML engine against which the MIA is launched over the air. For the PHY-layer authentication of potentially massive number of heterogeneous users (e.g., Internet of Things (IoT) devices), a service provider (e.g., gNodeB in 5G applications such as network slicing) can use such a classifier to classify users as authorized or not based on the RF fingerprints in the received signals (reflecting the inherent characteristics of the user's RF transceiver along with channel effects) and then admit communication requests of authorized users (e.g., it can be potentially implemented as xApps in the near Real-Time RAN Intelligent Controller (Near-RT RIC) of ORAN). The adversary can sense the spectrum to observe the behavior of a target classifier and then launch the MIA to determine whether a data sample (a wireless signal) of interest is in the training data of the target classifier or not. This attack reveals whether a wireless signal classifier is trained against a particular waveform, radio device, or channel environment. This private information leakage can be further exploited by the adversary to launch other attacks. For example, the adversary can spoof signals similar to the ones from authorized users using the same type of radio device and waveform and under a similar spectrum environment. This way, the adversary can bypass the signal classifier trained for PHY layer authentication, and can gain network access or prevent access of other users by occupying communication resources. The wireless systems pose unique challenges in data collection and design for the MIA that are different from other data domains such as computer vision. While an eavesdropper can observe a transmitted signal over the air, its

received signal is different from (but potentially correlated with) the signal received by the target signal classifier due to different channel characteristics. Therefore, the data collected by the adversary is inherently different from the data input to the target signal classifier. For RF fingerprinting, the service provider runs its DL classifier to determine whether the received signal is from an authorized user or not. The input of the DL classifier is the I/Q data and the underlying decision process of user classification is based on the RF fingerprint of the user that depends on the radio device, waveform and channel characteristics of that particular user. In this paper, we consider a black-box MIA, where the adversary does not know the target classifier. Even if the adversary knows this classifier (namely, the underlying DNN model), it may not use it to identify whether a signal is from an authorized user or not, since the signal received at the adversary is different from the signal received at the service provider. To overcome these challenges, the adversary builds a surrogate classifier by using the overheard signals as the input. With this surrogate classifier, the adversary can launch the MIA to determine for its received signal, whether the corresponding signal received at the service provider has been used in the training data or not.

We set up test scenarios of one service provider (such as a g Node B in 5G or beyond applications) and some authorized users (such as user equipments (UEs) in terms of IoT devices). Signals of each user are transmitted over the air and thus are changed by both channel and device-specific phase shift and transmit power effects. The target DL classifier can reliably classify users (with close to 100% under various settings). On the other hand, an adversary observes spectrum data and classification results (by observing whether a user is accepted for communications) to build a surrogate classifier to classify its received signals. The adversary then launches the MIA to infer whether for a signal received at the adversary, its corresponding signal received at the service provider is a member of the training data or not. We consider two settings: (i) non-member signals (signals that are not in the training data set) can be generated by the same radio devices that generate member signals (signals that are in the training data set), or (ii) non-member signals are generated by other radio devices. In the first setting, the accuracy of the MIA reaches 88.62% when the signal-to-noise-ratio (SNR) is high (at 10 dB) while the accuracy of the MIA is 77.01% when the SNR is low (at 3 dB). In the second setting, there are some radio devices that generate signals as training data. These signals cover both strong and weak signals. There is also another device that generates signals as non-member data to test the MIA performance. The accuracy of the MIA reaches 97.88%. Since wireless channels are random, they add uncertainties on received signals. Therefore, we study the impact of noisy variations in received signals by changing the question on whether a particular received signal is in training data to the question on whether some its noisy variations are in training data. For that purpose, we generate multiple samples with different levels of noisy variations and use either the average or maximum score in the MIA when evaluating the attack success. If we use the

average score, the accuracy of the MIA decreases with the level of noisy variations. If we use the maximum score, the accuracy of the MIA on member samples (authorized users) increases while the accuracy of the MIA on non-member samples (unauthorized users) decreases with the level of noisy variations.

We then develop a proactive defense scheme for the MIA. The service provider first needs to build a shadow MIA model. Then, it applies the defense using this shadow MIA model. For that purpose, perturbations (some controlled noise) are added in the classification process such that (i) there is no change made on classification results and (ii) the MIA in the presence of defense achieves low accuracy. We formulate this defense as an optimization problem and modify it to an unconstrained optimization by changing of variables and using loss function to remove constraints. We then apply gradient search to find the optimal perturbation. We show that this defense scheme effectively protects against the MIA launched by the adversary and reduces the accuracy from 97.88% to 50%.

2. LITERATURE SURVEY

"Deep learning for wireless communications,"

Wireless communications are envisioned to bring about dramatic changes in the future, with a variety of emerging applications, such as virtual reality, Internet of Things, and so on, becoming a reality. However, these compelling applications have imposed many new challenges, including unknown channel models, low-latency requirement in large-scale super-dense networks, and so on. The amazing success of deep learning in various fields, particularly in computer science, has recently stimulated increasing interest in applying it to address those challenges. Hence, in this review, a pair of dominant methodologies of using DL for wireless communications are investigated. The first one is DL-based architecture design, which breaks the classical model-based block design rule of wireless communications in the past decades. The second one is DL-based algorithm design, which will be illustrated by several examples in a series of typical techniques conceived for 5G and beyond. Their principles, key features, and performance gains will be discussed. Open problems and future research opportunities will also be pointed out, highlighting the interplay between DL and wireless communications. We expect that this review can stimulate more novel ideas and exciting contributions for intelligent wireless communications.

"When wireless security meets machine learning: Motivation, challenges, and research directions,"

Wireless systems are vulnerable to various attacks such as jamming and eavesdropping due to the shared and broadcast nature of wireless medium. To support both attack and defense strategies, machine learning (ML) provides automated means to learn from and adapt to wireless communication characteristics that are hard to capture by hand-crafted features and models. This article discusses motivation, background, and scope of research efforts that bridge ML and wireless security. Motivated by research directions surveyed in the context of ML for wireless security, ML-based attack and

defense solutions and emerging adversarial ML techniques in the wireless domain are identified along with a roadmap to foster research efforts in bridging ML and wireless security.

"Adversarial machine learning in wireless communications using RF data: A review,"

Machine learning (ML) provides effective means to learn from spectrum data and solve complex tasks involved in wireless communications. Supported by recent advances in computational resources and algorithmic designs, deep learning (DL) has found success in performing various wireless communication tasks such as signal recognition, spectrum sensing and waveform design. However, ML in general and DL in particular have been found vulnerable to manipulations thus giving rise to a field of study called adversarial machine learning (AML). Although AML has been extensively studied in other data domains such as computer vision and natural language processing, research for AML in the wireless communications domain is still in its early stage. This paper presents a comprehensive review of the latest research efforts focused on AML in wireless communications while accounting for the unique characteristics of wireless systems. First, the background of AML attacks on deep neural networks is discussed and a taxonomy of AML attack types is provided. Various methods of generating adversarial examples and attack mechanisms are also described. In addition, an holistic survey of existing research on AML attacks for various wireless communication problems as well as the corresponding defense mechanisms in the wireless domain are presented. Finally, as new attacks and defense techniques are developed, recent research trends and the overarching future outlook for AML in next-generation wireless communications are discussed.

3. EXISTING SYSTEM

An existing system presents channel-aware adversarial attacks against deep learning-based wireless signal classifiers. There is a transmitter that transmits signals with different modulation types. A deep neural network is used at each receiver to classify its over-the-air received signals to modulation types. In the meantime, an adversary transmits an adversarial perturbation (subject to a power budget) to fool receivers into making errors in classifying signals that are received as superpositions of transmitted signals and adversarial perturbations.

First, these evasion attacks are shown to fail when channels are not considered in designing adversarial perturbations. Then, realistic attacks are presented by considering channel effects from the adversary to each receiver. After showing that a channel-aware attack is selective (i.e., it affects only the receiver whose channel is considered in the perturbation design), a broadcast adversarial attack is presented by crafting a common adversarial perturbation to simultaneously fool classifiers at different receivers.

The major vulnerability of modulation classifiers to over-the-air adversarial attacks is shown by accounting for different levels of

information available about the channel, the transmitter input, and the classifier model. Finally, a certified defense based on randomized smoothing that augments training data with noise is introduced to make the modulation classifier robust to adversarial perturbations

DISADVANTAGES

The system is not implemented Membership Inference Attack(MIA) against datasets which leads less security.

In conjunction with security threats, an emerging concern on ML-based solutions is not privacy, namely the potential leakage of information from the ML models to the adversaries.

4. PROPOSED SYSTEM

1) In this paper, we present the first MIA that is launched against a wireless classifier over the air to infer about training data and leak private information on waveform, device, and channel characteristics.

2) We consider two settings for the MIA: (i) the MIA should be able to identify signals from the same radio device as member and non-member, and (ii) nonmember signals are generated by different radio devices.

3) We extend the MIA such that it is launched by using not only received signals but also their noisy variations by accounting for channel variations.

4) We show through detailed numerical results that the success of the MIA is high, i.e., the MIA can infer the training data membership of the wireless signal classifier with high accuracy.

5) We present a defense scheme to protect wireless signal classifiers from the MIA and show that this defense can reduce the accuracy of the MIA significantly.

ADVANTAGES

- The goal of the MIA is to identify data samples that have been used to train a ML classifier (as studied in computer vision and other data domains.

- The proposed system developed a proactive defense scheme for the MIA. The service provider first needs to build a shadow MIA model. Then, it applies the defense using this shadow MIA model.

5. SYSTEM ARCHITECTURE

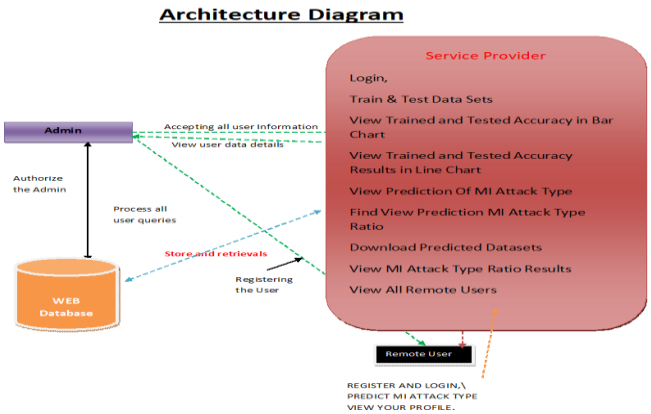

Fig 1: System architecture

The attacker needs access to the target model's predictions as well as some knowledge about the dataset used to train the model. The attacker then trains an attack model using the predictions of the shadow model as features and labels indicating whether each prediction came from the

training dataset or not. Finally, the attacker evaluates the attack model's performance on distinguishing between samples in the training dataset and those not in the training dataset. They may use metrics such as accuracy, precision, recall, and F1 score to assess the effectiveness of the attack.

6. IMPLEMENTATION

Modules

Service Provider:

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results in Line Chart, View Prediction Of MI Attack Type, Find View Prediction MI Attack Type Ratio, Download Predicted Datasets, View MI Attack Type Ratio Results, View All Remote Users.

View and Authorize Users:

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User:

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like PREDICT MI ATTACK TYPE, VIEW YOUR PROFILE.

7. IMPLEMENTATION RESULT

A membership inference attack is a type of privacy attack where an adversary attempts to determine whether a particular data point was used to train a machine learning model. These attacks rely on analyzing the model's predictions to infer membership of a given data point in the training dataset.

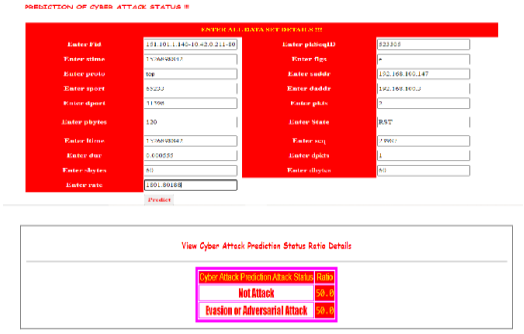

Fig 2. Implementation result

Implementing such an attack involves several steps, including training a shadow model to mimic the target model's behavior, generating synthetic data points, and comparing the target model's predictions on these data points with the shadow model's predictions. If there is a significant deviation between the two, it can indicate that the data point is a member of the training dataset. Evaluating the success of the attack typically involves measuring the accuracy of membership inference against a dataset with known membership labels.

8. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, we studied the MIA as a novel privacy threat against ML-based wireless applications. The target application is a DL-based classifier to identify authorized users by their RF fingerprint. An example use case for this attack is PHY-layer user authentication in 5G or IoT systems. The

input of this model consists of the received power and the phase shift. An adversary launches the MIA to infer whether signals of interest have been used to train this wireless signal classifier or not. In this attack, the adversary needs to collect signals and their classification results by observing the spectrum. Then, it can build a surrogate classifier namely a functionally equivalent classifier as the target classifier at the intended receiver, e.g., a service provider. We showed that the surrogate classifier can be reliably built by the adversary under various settings. Then, the adversary launches the MIA to identify whether for a received signal, its corresponding signal received at the service provider is in the training data or not

In the first setting where non-member signals can be generated by the same devices, the MIA accuracy is 88.62% for strong signals and 77.01% for weak signals. We studied the case that the member inference is investigated not only for received signals but also their noisy variations due to random channel effects. If the average score is used to predict the membership inference for original signals and their noisy variations, the accuracy of the MIA decreases with the level of noisy variations. On the other hand, if the maximum score is used, the accuracy on member samples increases while the accuracy on non-member samples decreases. In the second setting where non-member signals are generated by different devices, the MIA achieves better performance (97.88% accuracy). All these results indicate the MIA as a genuine threat for wireless privacy and show how the MIA can be effectively launched to infer private information from ML-based wireless systems over the air.

We further developed a defense scheme at the service provider that adds carefully crafted perturbations in the classification process such that there is no change on classification result but the MIA cannot work well. For the first setting, the MIA accuracy is originally not high and it is reduced by the defense only to a small extent (about 5%) while the defense is highly effective for the second setting and reduces the MIA accuracy to 50%.

REFERENCES

[1] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, “Over-the-Air Membership Inference Attacks as Privacy Threats for Deep Learning-based Wireless Signal Classifiers,” ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec) Workshop on Wireless Security and Machine Learning (WiseML), 2020.

[2] T. Erpek, T. O’Shea, Y. E. Sagduyu, Y. Shi, and T. C. Clancy, “Deep Learning for Wireless Communications” in Development and Analysis of Deep Learning Architectures, Springer, 2020

[3] Y. E. Sagduyu, Y. Shi, T. Erpek, W. Headley, B.Flowers, G. Stantchev, and Z. Lu, “When Wireless Security Meets Machine Learning: Motivation, Challenges, and Research Directions,” arXiv preprint arXiv:2001.08883, 2020/

[4] D. Adesina D, C. C. Hsieh, Y. E. Sagduyu, and L. Qian, “Adversarial Machine Learning in Wireless Communications using RF Data: A Review,” arXiv preprint arXiv:2012.14392, 2020.

[5] 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,” IEEE International Conference on Communications

(ICC) Workshop on Promises and Challenges of Machine Learning in Communication Networks, 2018.

[6] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep Learning for Launching and Mitigating Wireless Jamming Attacks," IEEE Transactions on Cognitive Communications and Networking, Mar. 2019.

[7] Y. Shi, Y. E. Sagduyu, T. Erpek, and M. C. Gursoy, "How to Attack and Defend 5G Radio Access Network Slicing with Reinforcement Learning," arXiv preprint arXiv:2101.05768, 2021. [8] M. Sadeghi and E. G. Larsson, "Adversarial Attacks on Deep-learning based Radio Signal Classification," IEEE Communications Letters, Feb. 2019.

[9] M. Sadeghi and E. G. Larsson, "Physical Adversarial Attacks Against End-to-end Autoencoder Communication Systems," IEEE Communications Letters, May 2019.

[10] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-Air Adversarial Attacks on Deep Learning Based Modulation Classifier over Wireless Channels," Conference on Information Sciences and Systems (CISS), 2020.

[11] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-Aware Adversarial Attacks Against Deep Learning-Based Wireless Signal Classifiers," arXiv preprint arXiv:2005.05321.

[12] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of Adversarial Attacks in DNN based Modulation Recognition," IEEE INFOCOM, 2020.

[13] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Adversarial Attacks with Multiple Antennas against Deep Learningbased Modulation Classifiers," IEEE Global Communications Conference (GLOBECOM), 2020.

[14] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, "Channel Effects on Surrogate Models of Adversarial Attacks against Wireless Signal Classifiers," IEEE International Conference on Communications (ICC), 2021.

[15] B. Manoj, M. Sadeghi, and E. G. Larsson, "Adversarial Attacks on Deep Learning based Power Allocation in a Massive MIMO Network," arXiv preprint arXiv:2101.12090, 2021.

[16] B. Kim, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial Attacks on Deep Learning Based mmWave Beam Prediction in 5G and Beyond," IEEE Statistical Signal Processing Workshop, 2021.

[17] R. Sahay, C. G. Brinton, and D. J. Love, "Ensemble-based Wireless Receiver Architecture for Mitigating Adversarial Interference in Automatic Modulation Classification," arXiv preprint arXiv:2104.03494, 2021.

[18] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust Adversarial Attacks Against DNN-Based Wireless Communication Systems," arXiv preprint arXiv:2102.00918, 2021.

[19] J. Yi, Jinho and A. El Gamal, "Gradient-based Adversarial Deep Modulation Classification with Data-driven Subsampling, arXiv preprint arXiv:2104.06375, 2021.

[20] Y. E. Sagduyu, Y. Shi, and T. Erpek, "IoT Network Security from the Perspective of Adversarial Deep Learning," IEEE International Conference on Sensing, Communication and Networking (SECON) Workshop on Machine Learning for Communication and Networking in IoT, 2019.