Mitigating Threats in PHY-Layer Authentication: A Proactive Defense Against Membership Inference Attacks in Wireless Signal Classifiers

D. Madhuri[1], V. Nikitha Reddy[1], M. Keerthi Reddy[1], V. N. L. N. Murthy[1], Saroja Kumar Rout[1(✉)], and Bijaya Kumar Sethi[2]

[1] Department of Information Technology, Vardhaman College of Engineering (Autonomous), Hyderabad, India
keerthireddymugala@gmail.com, rout_sarojkumar@yahoo.co.in

[2] Department of Computer Science and Engineering (Data Science), Vardhaman College of Engineering (Autonomous), Hyderabad, India

Abstract. In a wireless signal classifier utilized for PHY-layer authentication, a membership inference attack is demonstrated as an adversarial machine learning method. Waveform, channel, and device attributes are among the private information that needs to be retrieved. There is a difficulty since varying channel conditions produce varying received signals and RF fingerprints. The attacker constructs a surrogate classifier by examining the spectrum in order to circumvent this issue. Subsequently, we employ this surrogate model to conduct a black-box Membership Inference Attack (MIA) on the designated classifier. Our findings reveal that the adversary can effectively discern signals and potentially extract radio and channel information utilized in training the target classifier. To address this potential threat, we have implemented a proactive defense strategy. In order to fool the opponent, this involves creating a shadow MIA model. In order to reduce the MIA's accuracy and prevent data from the wireless signal classifier from leaking, faults are intended to be introduced. This scenario holds significance as it sheds insight on potential vulnerabilities in wireless signal classifiers, particularly with regard to PHY-layer authentication. In order to enhance wireless communication system security, the proactive defense strategy highlights how important it is to anticipate and prevent adversarial machine learning attacks.

Keywords: Adversarial machine learning · Membership Inference attack · Privacy · Wireless signal classification · Defense

1 Introduction

The Wireless networks are dynamic, and machine learning (ML), which has shown to be a successful approach to solving complicated issues in wireless communications affected by factors including channel conditions, interference, and traffic effects, has successfully addressed this. In particular, recent advances in computation and algorithms have

X. Cheng (Ed): BROADNETS 2024, LNICST 601, pp. 14–24, 2025.
https://doi.org/10.1007/978-3-031-81168-5_2

strengthened deep learning (DL), making it possible for it to efficiently capture high-dimensional representations of information on the spectrum [1]. The effectiveness of deep learning is demonstrated by a wide range of wireless communication applications, such as spectrum allocation, signal classification, waveform design, and spectrum sensing. However, there are unique security issues when ML/DL is integrated into wireless networks.

Due to adversarial machine learning (AML), attacks against ML/DL engines in wireless systems have expanded in variety. These attacks include those that facilitate clandestine communications, Trojan, spoofing, evasion (adversarial), and inference (exploratory). AML-based attacks function with small spectrum footprints, which makes detection more challenging, in contrast to more traditional wireless assaults such data transmission jamming [2]. Alongside security problems, privacy concerns are increasingly more common in machine learning (ML) solutions, particularly given the potential for adversaries to gain information from ML models. For instance, in a model inversion attack, adversaries have access to a machine learning model and particular private data, and they then use the model's inputs and outputs to deduce further private data. Another well-studied privacy assault is the membership inference attack (MIA), which has been applied to computer vision, healthcare, and commerce, among other data areas. MIA's goal is to ascertain if a certain data sample was part of the training set or not. MIA has been recognized as a significant privacy risk in the field of computer vision and related sectors, even if its potential application to the wireless sector is still unresolved (Fig. 1).

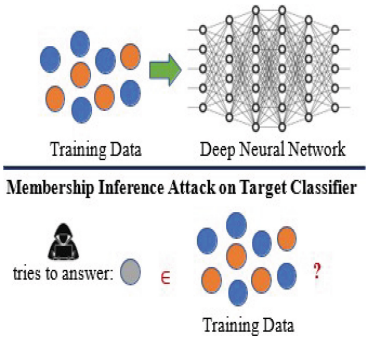

Fig. 1. Membership Inference Attack

Adversaries have particular chances to drop off over-the-air MIAs against wireless signal classifiers and eavesdrop on wireless signals due to the broadcast and shared aspect of the wireless channel. Since the target signal classifier's ML/DL model is trained using the waveform, channel, and basic radio device factors, they can therefore deduce information about these parts [5]. The present study describes the first-ever wireless application of Membership Inference Attack (MIA). Our aim is to introduce a goal for the aerial MIA machine learning (ML) technique, designed to function as a wireless signal classifier with enhancements from a deep neural network (DNN). This classifier serves the purpose of physical (PHY)-layer authentication for a wide range of diverse users, including Internet of Things (IoT) devices, such as those found in a gNodeB within

5G applications or network slicing scenarios. Service providers can utilize this classifier to authenticate users efficiently [6]. Using the Radio Frequency (RF) fingerprints found in received signals, this classifier helps determine if a user is authorized or unauthorized. These fingerprints show channel effects along with intrinsic characteristics of the user's RF transceiver. Then, requests for communication from approved users can be accommodated. This classifier could be used as an x App by the Open Radio Access Network's (ORAN) Near-Real-Time RAN Intelligent Controller (Near-RT RIC) [7]. Here, the adversary uses the MIA to ascertain whether the particular radio signal of interest is present in the desired classifier's training set after observing the target classifier's activity on the spectrum. The attack exposes whether particular waveform patterns, radio devices, or channel environments were utilized in training the wireless signal classifier. This breach of security could lead to further attacks exploiting the compromised personal data. For instance, adversaries could simulate the transmission of signals. This form of attack discloses whether a wireless signal classifier was trained under specific conditions, including waveform characteristics, device types, or radio environment settings [8].

The attacker could leverage the compromised personal information to carry out other attacks. Using a similar radio device type and waveform in a similar spectrum environment, for example, the attacker can fake signals that look to be coming from enabled users. Using signal classifiers designed for PHY-layer authentication, the adversary can steal communication resources thereby gaining network access or hindering other users' access. Wireless systems present distinct issues for MIA analysis and design than other data domains like computer vision. Even when an eavesdropper can see a provided signal over the air, the received signal differs from the signal received by the target signal classifier [10]. This leads to the data that the attacker collects and the data that the target signal classifier receives to fundamentally diverge from one another. The service provider uses its DL classifier to identify if the signal it receives in the context of RF fingerprinting is from an authorized user.

The DL classifier receives input in both phase and quadrature (I/Q) formats [11]. Based on the user's RF fingerprint, the categorization procedure takes into account the waveform, channel properties, and radio device. The adversary in a black-box MIA, or enemy unaware of the target classifier, is the focus of this work [12, 13]. Given the signal discrepancies that the adversary and the service provider receive, the adversary might not be able to use information about this classifier—that is, the underlying DNN model—to establish whether a signal is from an authorized user or not. The attacker builds a replacement classifier utilizing overheard signals as input to get past these barriers [14, 15]. The adversary can perform the MIA and ascertain whether the signal it received at the matching signal of the service provider was part of the signal's training data by utilizing this surrogate classifier.

2 Related Works

Recent research has increasingly focused on the intricate interplay between privacy invasion attacks directed at machine learning and deep learning models and the concept of differential privacy [1]. One such study by Rahman et al. delved into this relationship,

with a particular emphasis on models built on neural networks. In order to investigate the trade-off between privacy and utility in the context of membership inference attacks, their study involved modifying the privacy budget [2]. Conversely, Wang et al. focused on developing a differentially private regression model in order to counteract model inversion attacks in regression models. To maintain utility and guarantee differential privacy, their method made use of functional mechanisms. Zhang et al. conducted research on obfuscation techniques, which involve the introduction of noise into input datasets for the purpose of training machine learning models.

Their results showed that, in comparison to non-private scenarios, data reconstructed by model inversion attacks showed increased blurriness when applied to obfuscated models [3, 4]. Park et al. continued this line of inquiry by examining the connection between model inversion attacks and differential privacy, specifically in the context of face recognition software based on neural network models [5]. With regard to model inversion attacks, they carefully examined the trade-offs between privacy and utility under various privacy guarantee levels [8, 9]. Furthermore, Jayaraman and Evans investigated the relationship between attacks that violate privacy and differential privacy definitions, focusing on membership inference and attribute inference attacks. Neural network and logistic regression models were used in their investigation. Their results showed that, in comparison to non-private scenarios, data reconstructed by model inversion attacks showed increased blurriness when applied to obfuscated models [3, 4].

Normalized Least Mean Square Algorithm

An adaptation process is always influenced by the stability, convergence time, and volatility of the LMS as well as the step size. An efficient method of overcoming the update step size is to normalize the input signal's variance, $\sigma_u(t)^2$. Therefore, the following equation shows the weight update formula:

$$w(t+1) = w(t) + \frac{\mu}{N\sigma_u(t)^2} x(t)e^*(t) \tag{1}$$

In consequence, the LMS algorithm's performance asymptotically does not depend on the number of taps N. This has a significant effect on convergence rate. More taps result in poorer convergence rates. Park et al. continued this line of inquiry by examining the connection between model inversion attacks and differential privacy, specifically in the context of face recognition software based on neural network models [5]. With regard to model inversion attacks, they carefully examined the trade-offs between privacy and utility under various privacy guarantee levels [8, 9]. Furthermore, Jayaraman and Evans investigated the relationship between attacks that violate privacy and differential privacy definitions, focusing on membership inference and attribute inference attacks. Neural network and logistic regression models were used in their investigation 2019) expanded on this idea by addressing vulnerabilities in deep learning models and suggesting countermeasures in the Proceedings of the 2019 IEEE Symposium on Security and Privacy [14]. Numerous defense strategies have been invested, such as: Differential privacy involves making it more difficult for attackers to deduce membership by adding noise to the model's output or by using strategies like federated learning. Adversarial Training: To improve robustness, train the model as an adversarial task against membership inference attacks [15]. Regularization Techniques: Reducing overfitting and complicating

membership inference by using techniques like weight decay and dropout. Data augmentation involves adding artificial samples or perturbed data to training sets in order to make it harder for attackers to determine membership.

Ensemble Methods: To improve privacy and obstruct successful membership inference, combining multiple models or assembling [16, 17]. Private Aggregation of Teacher Ensembles (PATE): A method that protects privacy by training a student model with multiple teacher models, particularly in situations where sensitive data must be computed without dis-closing specific data points. The particular use case and threat model may influence the defense strategy selection [17, 18]. In order to protect sensing systems and deep learning models from membership inference attacks, researchers are constantly looking into new methods and improvements. Keeping up with the most recent findings in this area is essential for developing successful defense tactics.

3 Proposed Work

The study introduces the inaugural Membership Inference Attack (MIA) deployed over the air, aiming to infer training data and expose confidential details regarding waveform, device, and channel characteristics. Two MIA configurations are examined:

1. Various radio devices create nonmember signals, and
2. Both members and non-members' signals from the same radio device shall be distinguishable by the MIA.

The system extends the capabilities of Membership Inference Attack (MIA) by incorporating both original received signals and their corresponding noisy variations, taking into consideration the variations introduced by the channel. Extensive numerical analysis demonstrates the effectiveness of the MIA, indicating its ability to accurately deduce the membership of the training data used by the wireless signal classifier, as illustrated in Fig. 2.

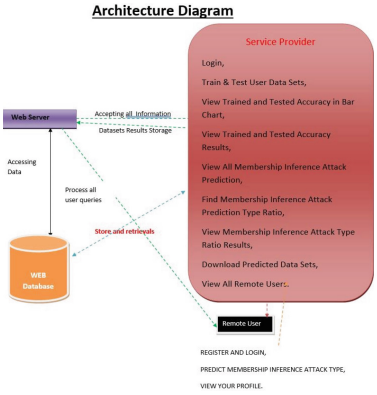

Fig. 2. Depicting the Architecture

The research work represents a defense approach to keep the MIA away from wireless signal classifiers and show how this defense can significantly reduce the accuracy of

the MIA. This work introduces the novel Membership Inference Attack (MIA). It is used to remotely attack a wireless classifier with the goal of gaining insights from the training set and disclosing personal data regarding the properties of the device, channel, and waveform. Two distinct situations for the MIA are examined: (i) testing the MIA's capacity to discriminate between members and non-member signals coming from the same radio device, and (ii) creating nonmember signals coming from other radio devices. Structures Create a diagram that takes into consideration the noisy changes of the channel fluctuations. Encompassing numerical findings demonstrate the notable success of the MIA as well as the accuracy with which it can identify which training data belongs to the wireless signal classifier. A defensive method is described to protect wireless signal classifiers against the MIA, showing how effective it is in significantly reducing the attack's accuracy.

In our test scenarios, a single service provider (e.g., a gN-odeB in 5G or beyond applications) and IoT user equipment (UEs) represent authorized consumers. Channel circumstances, device-specific phase shifts, and transmit power effects all impact each user's signals. Regarding user classification accuracy, the intended deep learning (DL) classifier consistently demonstrates great accuracy, nearly hitting 100% in many scenarios. Using the spectrum data and classification results, an adversary simultaneously constructs a surrogate classifier to categories the signals it receives. A Membership Inference Attack (MIA) is then used by the adversary to determine whether a signal received at the adversary is related to a member of the training data or not. We take into account two situations: The radio equipment that yields member signals can also generate non-member signals, or signals that are not part of the training dataset. Other radio equipment is the origin of non-member communications. As an example, the MIA accuracy in the first case is 77.01% at a low Signal-to-Noise Ratio (SNR) of 3 dB and 88.62% at a high SNR of 10 dB. One MIA produces training signals, and another radio device emits non-member signals in the second scenario.

We evaluate, taking into consideration the uncertainty induced by the stochastic nature of wireless networks, how noisy variations impact received signals. Finding out if the training data contains the received signal or any of its noisy variations is what this involves. Accuracy of the MIA decreases with increasing degree of noise variability when using the average score. The accuracy of the MIA, however, increases with the highest score when applied to member samples (authorized users) while decreasing with the number of noisy changes in non-member data (unauthorized users).

Noisy variations increase with the number of authorized users decreasing. In order to counter the MIA, we offer an active defense strategy. The service provider creates a shadow MIA model to carry out the defense strategy of employing controlled noise during the classification phase. This perturbation aims to maintain classification results while achieving poor accuracy for the MIA in the presence of defense. After a few variable adjustments and the application of a loss function to eliminate constraints, the defense strategy transitions from an optimization issue to an unconstrained optimization. Gradient search is then used to determine which perturbation is optimal. This defense strategy successfully thwarts the adversary-launched MIA, reducing accuracy from 97.88% to 50%.

4 Implementation

4.1 Navie Bayes

Naive Bayes: The Naive Bayes approach is a supervised learning technique based on the simple premise that a feature's presence or absence is independent of the presence or absence of any other feature in the class. This method has been shown to be reliable and effective, exhibiting comparable performance to other supervised learning techniques, despite its seemingly simple assumption. Its success is partly attributed by researchers to representation bias. Like linear discriminant analysis, logistic regression, or linear SVM, the Naive Bayes classifier is a linear classifier. Its simplicity in programming, straightforward parameter estimation, speedy learning on big databases, and passably good accuracy account for its popularity in the research domain. Revaluating learning outcomes is necessary for better comprehension and ease of implementation, though, due to its limited interpretability and difficulties in practice.

4.2 K Nearest Neighbour

K Nearest Neighbour is a simple yet effective classification algorithm utilizes a similarity metric to classify non-parametric and uses a lazy learning strategy, postponing learning until a test example is given entails classifying new data by determining K-nearest neighbors from the training set. As an example, the training dataset uses instances to determine learning and consists of the K-closest examples in feature space.

4.3 SVM

A discriminant machine learning technique in classification tasks looks for a discriminant function that, given an independent and identically distributed (ii d) training dataset, predicts labels for new instances. A data point is assigned to one of the several classes in the classification task by discriminant classification functions, as opposed to generative machine learning techniques, which call for the computation of conditional probability distributions. SVM, a discriminant technique, provides the same ideal hyperplane parameters every time it solves the convex optimization problem analytically. On the other hand, perceptron's and genetic algorithms (GAs) have the potential to produce distinct models with every training initialization. Perceptron's and GAs work to reduce error during training, which leads to a number of hyperplanes that satisfy the requirement (Fig. 3).

Fig. 3. Percentage of Attacks

4.4 Gradient Boosting

Gradient boosting is a machine learning technique composed primarily of weak prediction models, most frequently decision trees, that generates a prediction model in the form of an ensemble. It is used in regression and classification tasks [1, 2]. The algorithm that results from using decision trees as weak learners is called gradient-boosted trees, and it frequently performs better than random forests. Similar to other boosting techniques, the gradient-boosted trees model is developed step-by-step. However, it differs in that any differentiable loss function can be optimized, offering a more flexible and comprehensive capability.

5 Experimental Result

Accuracy: This parameter represents how well the attack model performs overall in identifying samples that belong to the training set and those that don't.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision: The fraction of samples classified as belonging to the training set that actually belong to the positive class is evaluated by precision.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall: Recall is a metric that expresses how many real positive samples the model properly identified in the in training set.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1-Score: This score provides an impartial assessment of the overall performance of the model by combining precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{5}$$

Figure 4 describes the Ml model is trained on the attack in top k vector dataset. Table 1 shows the different performance metrics of MIA attack. Figure 5 describes MIA ROC Curve CIFAR10. Figure 6 represents MIA ROC Curve CIFAR100.

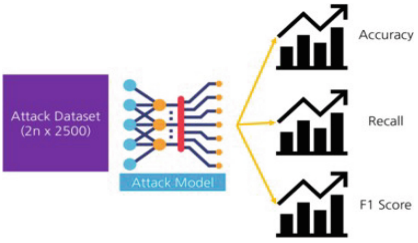

Fig. 4. Ml model is trained on the attack in top k vector dataset.

Table 1. Different performance metrics of MIA attack

MIA Attacks Metrices	Accuracy	Precision	Recall	F1 Score
CIFAR10	0.7761	0.7593	0.8071	0.7825
CIFAR100	0.9746	0.9627	0.9875	0.9749

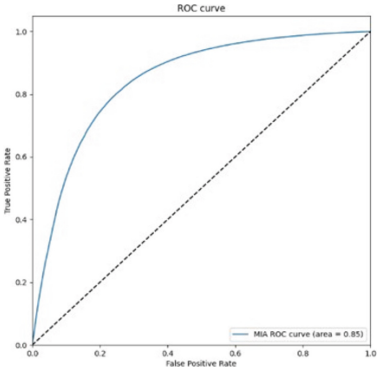

Fig. 5. MIA ROC Curve CIFAR10

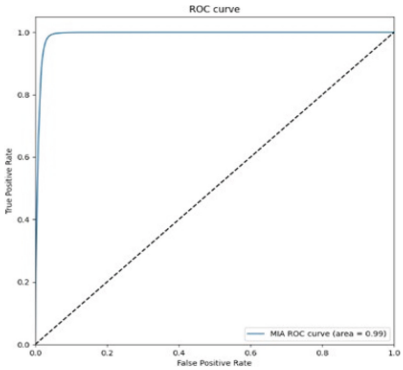

Fig. 6. MIA ROC Curve CIFAR100

6 Conclusions

Membership Inference Attacks (MIA), a fresh privacy problem in machine learning-driven wireless applications, were investigated in this paper. Our goal was to create a Deep Learning (DL) classifier that used radio frequency (RF) fingerprints to distinguish authorized users, which would be especially useful in 5G or Internet of Things (IoT) networks. As input features, the model makes use of phase shift and received power. The adversary uses MIA to scan for signals, build a surrogate classifier, and effectively identify whether a received signal matches the training set, attaining 77.01% accuracy for weak signals and 88.62% accuracy for strong signals increasing noisy variations reduce membership inference accuracy using average scores, particularly affecting non-member samples, where utilizing maximum scores results in decreased accuracy. In scenarios where non-member signals are from distinct devices, the Membership Inference Attack (MIA) excels, achieving 97.88% accuracy. This underscores MIA's threat to wireless privacy. A defensive tactic, involving deliberate perturbations in the categorization process, is devised. While reducing MIA accuracy by around 5% in the first scenario, it performs exceptionally in the second, lowering accuracy to 50%.

References

1. Davaslioglu, K., Soltani, S., Erpek, T., Sagduyu, Y.E.: DeepWiFi: cognitive WiFi with deep learning. IEEE Trans. Mob. Comput. (2021)
2. Hui, B., Yang, Y., Yuan, H., Burlina, P., Gong, N.Z., Cao, Y.: Practical blind membership inference attack via differential comparisons. In: Network and Distributed System Security Symposium (NDSS) (2021)
3. Song, L., Mittal, P.: Systematic evaluation of privacy risks of machine learning models. In: USENIX Security Symposium (2021)
4. Kim, B., Sagduyu, Y.E., Erpek, T., Davaslioglu, K., Ulukus, S.: Channel effects on surrogate models of adversarial attacks against wireless signal classifiers. In: IEEE International Conference on Communications (ICC) (2021)
5. Yi, J., El Gamal, A.: Gradient-based adversarial deep modulation classification with data-driven subsampling. arXiv preprint arXiv:2104.06375 (2021)

6. Erpek, T., Sagduyu, Y.E., Shi, Y.: Deep learning for launching and mitigating wireless jamming attacks. IEEE Trans. Cogn. Commun. Netw. (2019)
7. Shi, Y., Sagduyu, Y.E., Erpek, T., Gursoy, M.C.: How to attack and defend 5g radio access network slicing with reinforcement learning. arXiv preprint arXiv:2101.05768 (2021)
8. Sagduyu, Y.E., et al.: When wireless security meets machine learning: motivation, challenges, and research directions. arXiv preprint arXiv:2001.08883 (2020)
9. Sadeghi, M., Larsson, E.G.: Physical adversarial attacks against end-to-end autoencoder communication systems. IEEE Commun. Lett. (2019)
10. Kim, B., Sagduyu, Y.E., Davaslioglu, K., Erpek, T., Ulukus, S.: Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels. In: Conference on Information Sciences and Systems (CISS) (2020)
11. Kim, B., Sagduyu, Y.E., Davaslioglu, K., Erpek, T., Ulukus, S.: Channel-aware adversarial attacks against deep learning-based wireless signal classifiers. arXiv preprint arXiv:2005.05321
12. Lin, Y., Zhao, H., Tu, Y., Mao, S., Dou, Z.: Threats of adversarial attacks in DNN based modulation recognition. In: IEEE INFOCOM (2020)
13. Kim, B., Sagduyu, Y.E., Davaslioglu, K., Erpek, T., Ulukus, S.: Adversarial attacks with multiple antennas against deep learning based modulation classifiers. In: IEEE Global Communications Conference (GLOBECOM) (2020)
14. Rout, S.K., Rath, A.K., Bhagabati, C.: Energy efficient and cost effective secure node localization with key management in wireless sensor networks. In: 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 515–520. IEEE (2016)
15. Manoj, B., Sadeghi, M., Larsson, E.G.: Adversarial attacks on deep learning based power allocation in a massive MIMO network. arXiv preprint arXiv:2101.12090 (2021)
16. Kim, B., Sagduyu, Y.E., Erpek, T., Ulukus, S.: Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond. In: IEEE Statistical Signal Processing Workshop (2021)
17. Sahay, R., Brinton, C.G., Love, D.J.: Ensemble-based wireless receiver architecture for mitigating adversarial interference in automatic modulation classification. arXiv preprint arXiv:2104.03494 (2021)
18. Bahramali, A., Nasr, M., Houmansadr, A., Goeckel, D., Towsley, D.: Robust adversarial attacks against DNN-based wireless communication systems. arXiv preprint arXiv:2102.00918 (2021)
19. Shi, Y., Davaslioglu, K., Sagduyu, Y.E.: Over-the-air membership inference attacks as privacy threats for deep learning-based wireless signal classifiers. In: ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec) Workshop on Wireless Security and Machine Learning (WiseML) (2020)
20. Chen, D., Yu, N., Zhang, Y., Fritz, M.: GAN-leaks: a taxonomy of membership inference attacks against generative models. In: ACM SIGSAC Conference on Computer and Communications Security (CCS) (2020)