International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

MEMBERSHIP INFERENCE ATTACK AND DEFENSE FOR WIRELESS SIGNAL CLASSIFIERS WITH DEEP LEARNING

L.C. Usha Maheswari[1], K. Samatha[2], K. Akhila[3]

[1] Assistant Professor, School of CSE,Malla Reddy Engineering College For Women(Autonomous Institution), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100

[2,3]UG Scholar, Department of IOT,Malla Reddy Engineering College for Women, (Autonomous Institution), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100

Email: maheswari07usha@gmail.com

ABSTRACT

Wireless Membership Inference Attack (MIA) is presented to leak private information from wireless signal classifiers. Machine learning (ML) provides a powerful means to classify wireless signals. For PHY-layer authentication. As an adversarial machine learning attack, MIA infers whether a signal of interest is used in the training data of a target classifier. This private information includes waveform, channel, and device characteristics, which, if leaked, can be exploited by an attacker to identify vulnerabilities in the underlying ML model (e.g., penetrating PHY-layer authentication). The challenge with wireless MIA is that the received signal, and therefore the RF fingerprint, differs between the attacker and the intended receiver due to mismatched channel conditions. Therefore, an attacker first observes the spectrum and builds a surrogate classifier, and then launches black-box MIA on this classifier. The MIA results show that the attacker can reliably infer the signals (and possibly radio and channel information) used to build the target classifier. Therefore, an active defense against MIA is developed by building a shadow MIA model and deceiving the attacker. This defense can reduce the accuracy of MIA and prevent information leakage from the radio signal classifier.

Keywords-Wireless Membership Inference Attack (MIA), PHY-layer Authentication, Radio Signal Classifier, Adversarial Machine Learning

I. INTRODUCTION

In recent years, machine learning (ML) and deep learning (DL) have emerged as powerful tools in wireless communications, enabling systems to adapt to and learn from dynamic environments. These techniques have been applied to tasks such as spectrum sensing, signal classification, spectrum allocation, and waveform design to provide solutions to complex challenges in wireless networks, such as interference and traffic management. However, when ML/DL is incorporated into wireless systems, there come unique security and privacy challenges that must be addressed. For instance, one of the dangers is the MIA in the wireless environment, which has gained considerable attention because it can exploit the privacy vulnerabilities in ML models. MIA is an AML attack, which enables the inference by an attacker regarding whether a given data sample is from the training set of some target

ML model. In this sense, in the domain of wireless communications, it affects the classifiers trained on wireless signals and may disclose some secret information such as waveforms, radio characteristics, or even channel conditions. This type of information leakage can expose the RF fingerprint of a device and compromise the security of systems that rely on PHY layer authentication, such as 5G and IoT environments. The challenge of successful MIA in wireless communications arises from inherent differences in signal characteristics due to channel mismatch between the attacker and the intended receiver. To overcome this, an attacker can use intercepted spectrum data to create a surrogate classifier, thus allowing them to carry out black-box MIA without having to access the internal model of the target classifier.This method allows an attacker to determine if a received signal is part of the training data. This information, in turn, reveals significant information about the radio, waveform, and channel environment used to train the classifier. In this project, we focus on wireless membership inference attacks in the context of deep neural network (DNN)-based signal classifiers used for PHY-level authentication. These classifiers are designed to identify users who are: A. IoT devices using their unique RF fingerprint. An attacker could eavesdrop on the wireless spectrum and create a surrogate model to infer whether a particular signal is part of the training data, bypassing the authentication process. This could lead to a serious security breach in which an attacker could spoof authorized signals and gain unauthorized access to the network. We further explore the influence of channel variation on the effectiveness of MIA and demonstrate how an attacker can leverage noisy signal fluctuations for improving the accuracy of an attack based on the condition of the channel. In response to this threat, we propose a defense mechanism that allows service providers to proactively deceive attackers using a shadow MIA model. This defense reduces the precision of MIA by adding a controlled noise to the classification in addition to preserving the confidentiality of the underlying training data.

This is the first time that we use MIA for wireless communication systems, identify the potential risks of privacy violation, and propose a novel defense method against this kind of attack. With rigorous tests, we show that we are able to dramatically decrease the accuracy of MIA and ensure the security and privacy of wireless systems against such adversarial threats.

II. RELATED WORK

1.MIA in Traditional Domains

In traditional domains like computer vision, healthcare, and natural language processing, it has been shown that MIA can leak sensitive information. Shokri et al. (2017) proposed the concept of MIA in the context of machine learning in their seminal work. In MIA, an attacker tries to infer whether a given data point was used in the training set of a model. Several other studies, since then, have targeted the defense mechanism against MIA by coming up with methods to decrease the chances of successful membership inference. Such methods include techniques of differential privacy, adversarial training and strategies for model obfuscation. For instance, Hayes et al. (2019) showed how an attacker could infer membership from a black box model in deep learning models. Their work provided a foundation for understanding the risks of membership inference in ML systems. Yeom et al. (2018) extended MIA to machine learning models and presented several methods to counter such attacks, including using models that do not store enough information for an attacker to successfully infer membership.

2. MIA in wireless communication systems

In the wireless domain, RF fingerprinting is widely used nowadays for authentication purposes, which applies the unique features of transmitted signals to identify devices and users. Wang et al. (2020) discussed vulnerability in wireless systems due to adversarial attacks by focusing on the way attackers can eavesdrop on and manipulate the wireless signal to bypass authentications. However, the majority of these studies have been based on evasion attacks (where the attacker has control over the signal in an attempt to deceive the classifier) and thus privacy concerns such as MIA are hardly explored in this context. In recent years, researchers have begun to apply adversarial machine learning (AML) techniques, including MIA, to wireless systems. Shi et al. (2022) conducted the first deep study on the application of MIA in wireless communications with a focus on the classifiers used for PHY-layer authentication in 5G and IoT systems. They demonstrated that attackers can leverage the sharing and transmission nature of wireless signals to perform MIA to extract vital information, such as the waveforms, devices, and channel characteristics used in training the classifier. Their work also introduced a novel methodology in which a surrogate classifier is used to infer membership, even though the attacker does not have access to the internal structure of the target classifier. This study highlights the importance of mitigating such attacks that can lead to the compromise of secure wireless systems.

3.Adversarial Machine Learning (AML) in Wireless Systems

AML has been used in the context of wireless communications to counter various attacks such as

Evasion attacks (where the attacker changes the signal to evade detection) and Poisoning attacks (where the attacker injects malicious data into the training set). Zhang et al. (2020) proposed several adversarial attacks on wireless ML classifiers, among which Trojan attacks are that an attacker manipulates the model in training to achieve a certain goal after its deployment. However, their work focuses more on the vulnerability of wireless systems toward evasion attacks and does not directly address privacy issues such as MIA. Wang et al. (2021) investigated evasion-based adversarial attacks on the wireless signal classification models, which attackers can manipulate the transmitted signal to evade the classifier trained by the model. Although such attacks pose a great threat to the integrity of the system, they are not attacking leakage of private information, which is the purpose of MIA.Thus, Shi et al. (2022) is the first study that comes up with a detailed study of MIA in wireless systems, especially on signal classifiers relying on deep learning for PHY-level authentication.

4. Counter measures against MIA in wireless systems

Several mitigation strategies have been proposed in the broader machine learning community to mitigate the risks associated with MIA. Shokri et al. in (2017) propose a defense mechanism called differential privacy whereby a model cannot disclose that the particular data sample was used in training.

In Yeom et al. work in 2018, they used some privacy-preserving techniques on the training data to obscure the membership status of the data. Recently, Carlini et al. (2020) introduced a potentially controversial training technique that improved the robustness of the model by adding noise or perturbations to the decision boundary of the model, reducing the success rate of membership inference attacks. For wireless systems,Shi et al. (2022) proposed an active defense mechanism against MIA by constructing a shadow MIA model, which introduces noise (controlled noise) into the signal classification process to prevent attackers from accurately inferring membership, thereby protecting the confidentiality of training data. This defense effectively reduces the success rate of MIA and demonstrates its potential to protect against adversarial attacks on wireless networks.

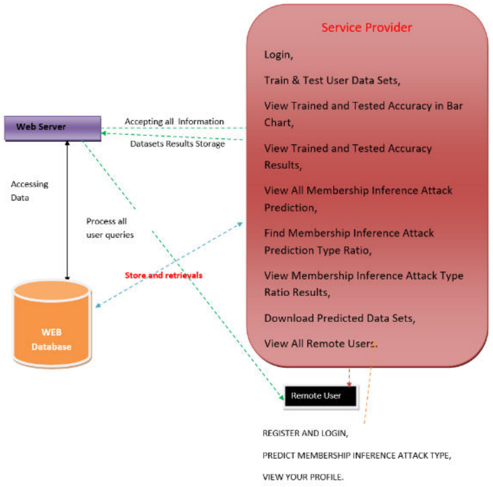

Service Provider
Login,
Train & Test User Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
View All Membership Inference Attack Prediction,
Find Membership Inference Attack Prediction Type Ratio,
View Membership Inference Attack Type Ratio Results,
Download Predicted Data Sets,
View All Remote Users.
Web Server
Accepting all Information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
PREDICT MEMBERSHIP INFERENCE ATTACK TYPE,
VIEW YOUR PROFILE.

Fig : System Architecture

III. IMPLEMENTATION

1.User Registration and Authentication :

To ensure that only authorized people can access the system and deny unauthorized access and guarantee security. **Remote Users:** A remote user first needs to apply by submitting their details and required information, such as email addresses and username, after which the admin shall examine his or her application to subsequently approve it.

Login Process: All the registered and approved customers must log in using an authentic username and password secured for login. This eliminates any unverified and unwanted users from accessing the function of the system.

2. Training and Testing the Model :

The main function of the system is training and testing a machine learning model to classify wireless signals with possible Membership Inference Attacks such as MIA.

Training the Model: The service provider or admin uploads the labeled datasets of wireless signals, which are often labeled as part of the training set or not. Using these, they train a machine learning model that can classify the wireless signals by using deep learning techniques.

Testing the Model: Once trained, the model should be tested on new, never-before-seen data to ensure its performance. This process of testing will determine the capability of the model to categorize signals correctly and prevent potential attacks such as MIA**.**

3. Displaying Model Results :

After training and testing the model, it is important that a service provider analyze its performance which is shown through reports and visualizations easy to understand.

Accuracy Visualization:

The system allows graphical representation, for example bar charts or graphs of the accuracy of training and testing the model. Therefore, the service provider would have an overview of whether the model is performing very well or requires improvement.

Membership Inference Attack (MIA) Results:

A very important test of the robustness of the system is the ease at which it can be attacked. For this module, the MIA results show how easily attackers can infer private training data from the outputs of the model.

4. Membership Inference Attack (MIA) Detection

The system shall ensure it's not prone to privacy attacks and especially to MIA when attackers aim to infer whether any given signal was used when training a model. This makes it check the defense of the system against this kind of attacks**.**

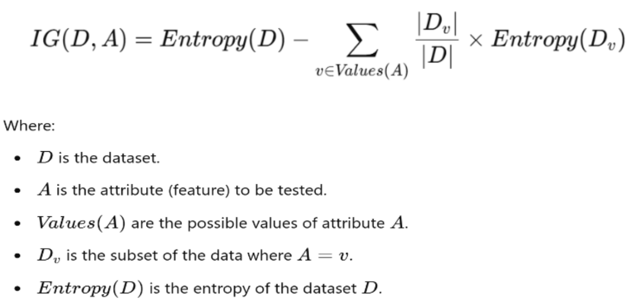
MIA Simulation: At this step, an attacker tries to infer whether a certain signal used by a remote user was included in the training data. The MIA helps assess the privacy risk of the model and how easily the attacker can access private information about users, their devices, or their channel characteristics.

Attack Type Prediction: The system will predict whether the MIA was successful, thus providing insight into how well the classifier resists the leakage of private data.

IV.ALGORITHMS USED

1.Decision Tree Classifiers :

Decision tree classifiers are used to make decisions based on the data. They create a flowchart-like tree where each branch represents a decision, and each leaf represents an outcome or class label. The decision tree helps to classify objects into categories based on their features.



$$IG(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \times Entropy(D_v)$$

Where:

- D is the dataset.
- A is the attribute (feature) to be tested.
- $Values(A)$ are the possible values of attribute A.
- D_v is the subset of the data where $A = v$.
- $Entropy(D)$ is the entropy of the dataset D.

2. Gradient Boosting :

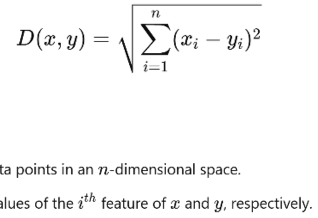
Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function

3.K-Nearest Neighbors (KNN) :

KNN is a simple and effective classification algorithm that looks at the closest data points to make predictions. When a new data point arrives, it checks its "neighbors" and classifies it based on the most common class among its neighbors.

The algorithm finds the nearest data points (neighbors) to the new data point based on distance (e.g., Euclidean distance). The new data is assigned the class that appears the most among its neighbors.

Euclidean Distance:



$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Where:

- x and y are two data points in an n-dimensional space.
- x_i and y_i are the values of the i^{th} feature of x and y, respectively.

4. Logistic Regression Classifiers :

Logistic regression is a statistical method for analyzing datasets in which the outcome is binary or categorical. It is used for binary classification tasks and predicts the probability of an instance belonging to a particular class (0 or 1).The algorithm uses a logistic function (also called the sigmoid function) to model the relationship between the dependent variable and the independent variables. The output of the logistic function is a probability, which is mapped to a binary class label (e.g., 0 or 1). Logistic regression is widely used for classification problems such as email spam detection, medical diagnosis (e.g., predicting the likelihood of a disease), and customer churn prediction.

5. Naive Bayes :

Naive Bayes is a probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions between the features. Despite its simplicity, it performs well for many types of classification problems. It is often used in text classification tasks, such as spam filtering, sentiment analysis, and document classification. It is also used in medical diagnosis, where the independence assumption can often be reasonable.

6. Random Forest : Random Forest is an ensemble learning methodology, which constructs a collection of decision trees during training and at testing time outputs the class label or regression value based upon the majority vote (for classification) or average for regression from all the trees.

The individual decision tree in a random forest is trained on a random subset of the data, applying bootstrapping, while using random feature selection at each split. The ensemble method reduces the variance of individual decision trees and therefore improves the accuracy of models.Due to its high accuracy and strength against overfitting, random forest is used in many applications, including medical predictions, fraud detection, credit scoring, and image classification.

7. **Support Vector Machine (SVM) :**

SVM is a type of supervised learning algorithm for the purpose of classification and regression tasks. It discovers a hyperplane with a maximum-margin gap that is the best hyperplane in differentiating classes in data. The kernel trick transforms the input space to a feature space that has higher dimension. A wide margin in separation for all points is achieved in such a space through this process. The objective is to maximize the margin between the support vectors (the closest points of each class).

SVM is also highly applied in text classification, image recognition, and other fields with high dimensional data where classification problems exist. In this sense, it becomes quite efficient in classification tasks which involve clear margins of separation.

V.RESULTS

Membership Inference Attack and Defense for Wireless Signal Classifiers w
Deep Learning

Adversarial machine learning,deep learning,membership inference attack,privacy,wireless signal classification,defense.
Login Using Your Account:
User Name
Password
Login

Fig:1:User Login

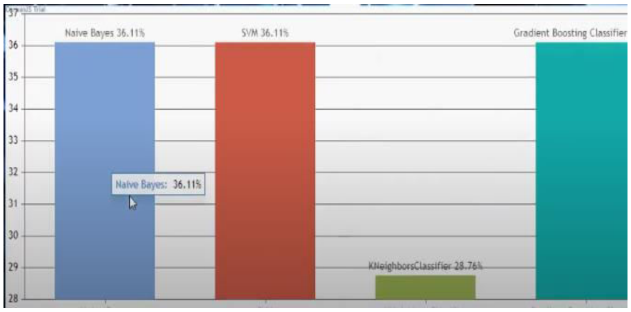


Fig:2:Accuracy Results

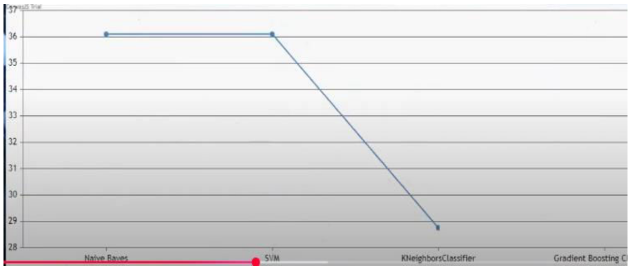


Fig:3:Accuracy Ratio

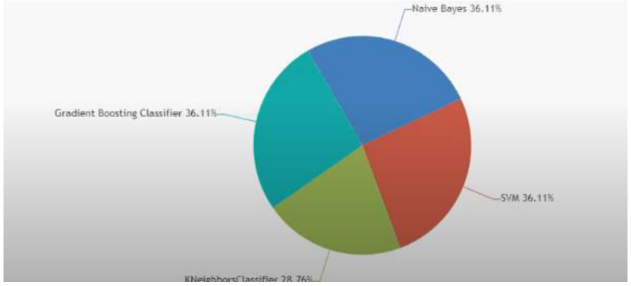


Fig:4:Pie chart of Algorithms Accuracy

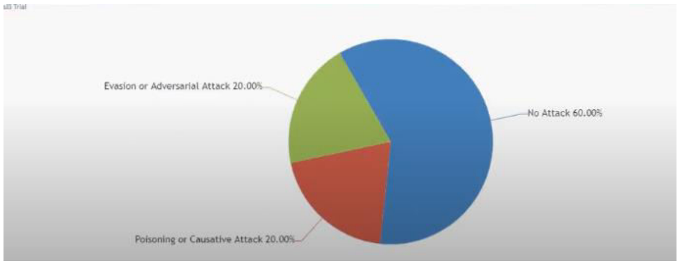


Fig:5:Pie chart Attackers

VI.CONCLUSION

In this paper, we took MIA as one of the emerging privacy threats for ML-based wireless applications. The application that is targeted is DL-based classification of authorized users with RF fingerprints. An example application for such an attack is the PHY-layer user authentication of 5G or IoT systems. The input to the model consists of received power and phase shift. An attacker can invoke MIA to check whether the signal of interest was used to train this wireless signal classifier. In this attack, the attacker has to observe the spectrum to collect signals and their classification results. Then, they can build a replacement classifier, i.e., a classifier that is functionally equivalent to the target classifier at the intended receiver. B. Service Provider. We demonstrated that the attacker can construct surrogates reliably under different settings. The attacker then launches MIA to determine, for any received signal, whether training data contains a corresponding signal received at the service provider. In the first setting, where non-member signals are generated from the same device, the MIA accuracy is 88:62% for strong signals and 77:01% for weak signals. We investigated the scenario where member inference inspection is not only over the received signal but also over noisy fluctuations originating from random channel effects. Using the mean value in predicting member inference over both the original signal and its noisy fluctuations degrades the accuracy of MIA with a large degree of noisy fluctuations. On the other hand, using the maximum value improves accuracy for member samples but decreases accuracy for non-member samples. In the second setting, where non-member signals are generated from different devices, MIA achieves better performance (97:88% accuracy). All these results demonstrate that MIA is indeed a real threat to wireless privacy and demonstrates how MIA can be effectively launched to infer private information over the air from ML-based wireless systems. We also developed a mitigation scheme at the service provider that injects carefully crafted perturbations into the classification process that do not change the classification results but make MIA perform worse. In the first setting, MIA's hit rate is not high to start and only very slightly reduced by the defense (around 5%). In the second setting, the defense is extremely effective and reduces MIA's hit rate to 50%.

REFERENCES

[1] B. Flowers, R. M. Buehrer, and W. C. Headley, “Evaluating Adversarial Evasion Attacks in the Context of Wireless Communications,” 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] M. DelVecchio, V. Arndorfer, and W. C. Headley, “Investigating a Spectral Deception Loss Metric for Training Machine Learning-based Evasion Attacks,” ACM Workshop on Wireless Security and Machine Learning, 2020.

[4] D. Adesina D, C. C. Hsieh, Y. E. Sagduyu, and L. Qian, Adversarial Machine Learning in Wireless Communications using RF Data 2020.

[5] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, “Over-the-Air Adversarial Attacks on Deep Learning Based Modulation Classifier over Wireless Channels,” 2020.

[6]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,” 2020.

[7] S. Kokalj-Filipovic and R. Miller, “Adversarial Examples in RF Deep Learning: Detection of the Attack and its Physical Robustness,” 2019.

[8] 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,” 2018.

[9] T. Erpek, Y. E. Sagduyu, and Y. Shi, “Deep Learning for Launching and Mitigating Wireless Jamming Attacks,”Mar. 2019.

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

[12] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, “Channel Effects

on Surrogate Models of Adversarial Attacks against Wireless Signal Classifiers," 2021.

[13] B. Kim, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial Attacks on

Deep Learning Based mmWave Beam Prediction in 5G and Beyond,"

2021.

[14] M. Sadeghi and E. G. Larsson, "Adversarial Attacks on Deep-learning based Radio Signal Classification," Feb. 2019.

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

[16] B. Kim, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial Attacks with Multiple Antennas against Deep Learning-based Modulation Classifiers," 2020.

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

[18] S. Kokalj-Filipovic, R. Miller, and J. Morman, "Targeted Adversarial Examples against RF Deep Classifiers,"2019.

[19] Y. Shi, Y. E. Sagduyu, T. Erpek, and M. C. Gursoy, "How to Attack and Defend 5G Radio Access Network Slicing with Reinforcement Learning," 2021.

[20] B. Flowers, M. DelVecchio, B. Flowers, A. J. Michaels, and W. C. Headley, "On the Limitations of Targeted Adversarial Evasion Attacks against Deep Learning Enabled Modulation Recognition," 2019.

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

[22] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Over-the-Air Membership Inference Attacks as Privacy Threats for Deep Learning-based Wireless Signal Classifiers,"2020.

[23] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, "Over-the-Air Adversarial Attacks on Deep Learning-Based Modulation Classifiers," 2020.

[24] Y. E. Sagduyu, Y. Shi, and T. Erpek, "IoT Network Security from the Perspective of Adversarial Deep Learning,"2019.

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