# STEALTHY ADVERSARIAL ATTACKS AGAINST AUTOMATED MODULATION CLASSIFICATION IN COGNITIVE RADIO

**Praveen Fernando and Jin Wei-Kocsis**

**Computer and Information Technology**

**Purdue University**

# AGENDA

- Introduction
- Problem formulation
- Proposed method
- Performance evaluations
- Conclusions

# AGENDA

- Introduction
- Problem formulation
- Proposed method
- Performance evaluations
- Conclusions

# INTRODUCTION

- In cognitive radio systems, wireless spectrum sensing plays a crucial role in identifying the state of the wireless environment, which leads to the effective utilization of scarce spectral resources for various application fields.
- As a critical part of wireless spectrum sensing, automatic modulation classification (AMC) is used to identify the modulation types of the received signals automatically.
- With the advances in sensing and computing technologies, artificial intelligence (AI), especially deep learning (DL), has been widely applied to enhance the effectiveness and timely response of AMC.

# INTRODUCTION

- In recent years, increasing evidence shows that carefully-crafted adversarial noise can introduce bounded subtle adversarial perturbation that can mislead learning models.
  - ✓Different adversarial attack methods have been proposed in computer vision and natural language processing domains, including the fast gradient sign method (FGSM), the fast gradient value (FGV) method, and basic iterative method (BIM).
- The mitigation of the DL performance caused by adversarial attacks has raised concerns about the trustworthiness of DL solutions in AMC.

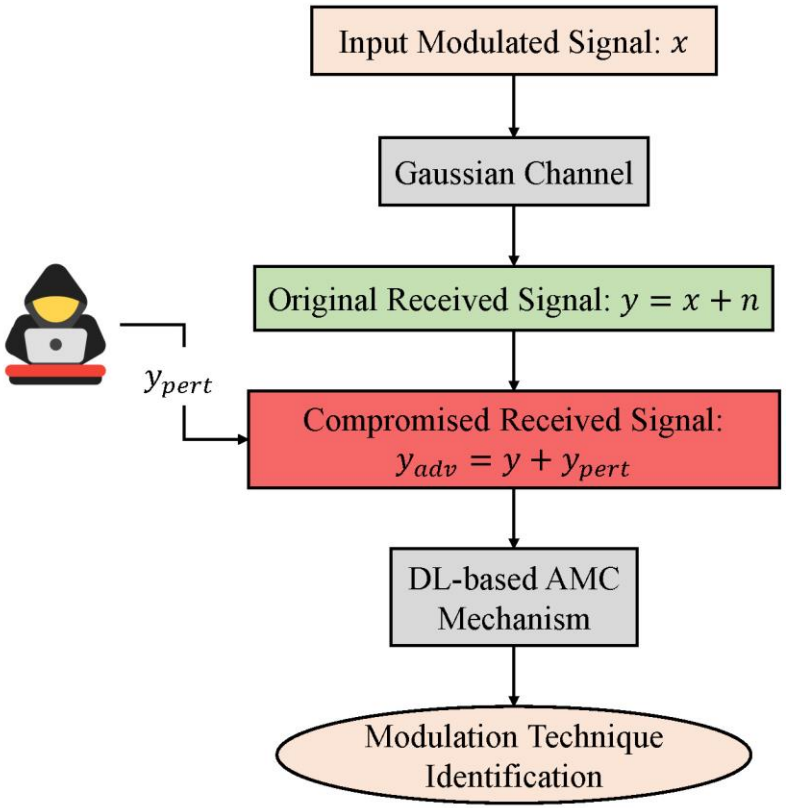
# INTRODUCTION

- It is crucial to sufficiently mitigate the adversarial perturbations in DL-powered AMC.
  - ✓To realize a successful mitigation strategy, it can be beneficial to first adopt an adversarial mindset and formulate threat models of practical perturbations in DL-powered AMC.
  - ✓There is still limited work focusing on generating adversarial attacks in the domain of wireless communications.
- In this paper, we propose an innovative and stealthy adversarial attack method that can practically compromise the DL-based decision-making of AMCs.

# AGENDA

- Introduction
- Problem formulation
- Proposed method
- Performance evaluations
- Conclusions

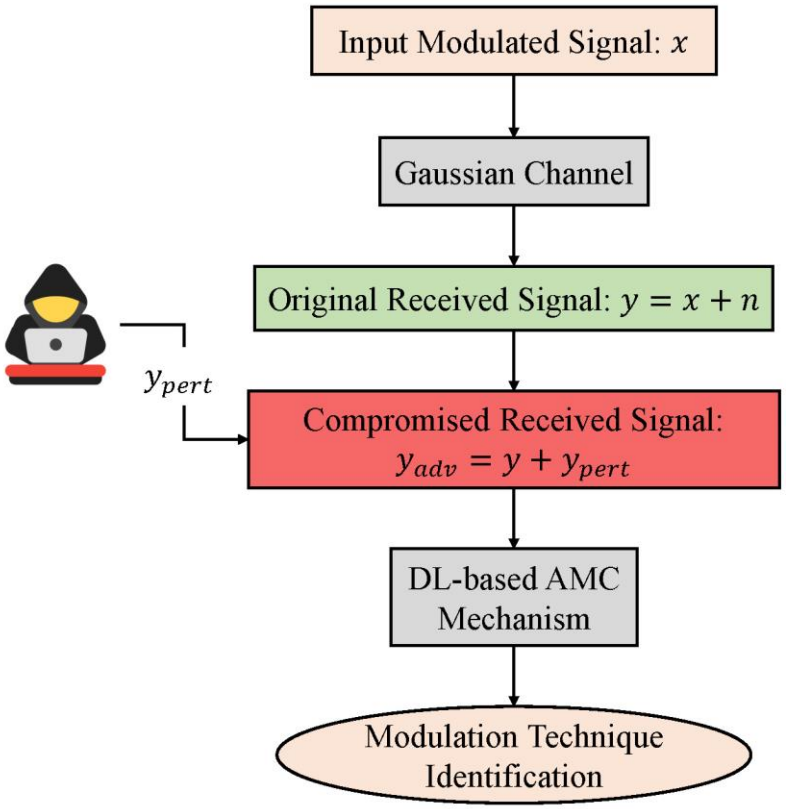
# PROBLEM FORMULATION




- We develop a practical and stealthy adversarial attack to compromise the automated decision-making process of a DL-powered AMC system.
  - ✓We assume that the AMC mechanism adopts two types of DL techniques, convolutional neural network (CNN) and gated recurrent unit (GRU), for modulation technique identification.

Fig. 1: Overview of the AMC system considered in our work.

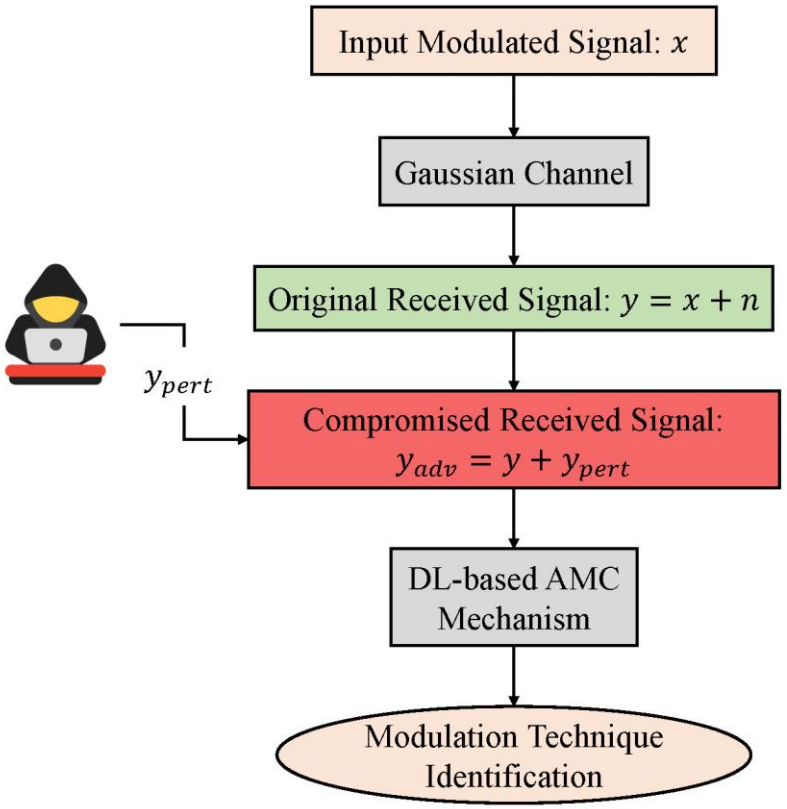
# PROBLEM FORMULATION




- We develop a practical and stealthy adversarial attack to compromise the automated decision-making process of a DL-powered AMC system.
  - ✓We assume that there are 24 potential different modulation techniques, including 32PSK, 16APSK, 32QAM, FM, GMSK, 32APSK, OQPSK, 8ASK, BPSK, 8PSK, AM-SSB-SC, 4ASK, 16PSK, 64APSK, 128QAM, 128APSK, AM-DSB-SC, AM-SSB-WC, 64QAM, QPSK, 256QAM, AM-DSB-WC, OOK, and 16QAM.

Fig. 1: Overview of the AMC system considered in our work.

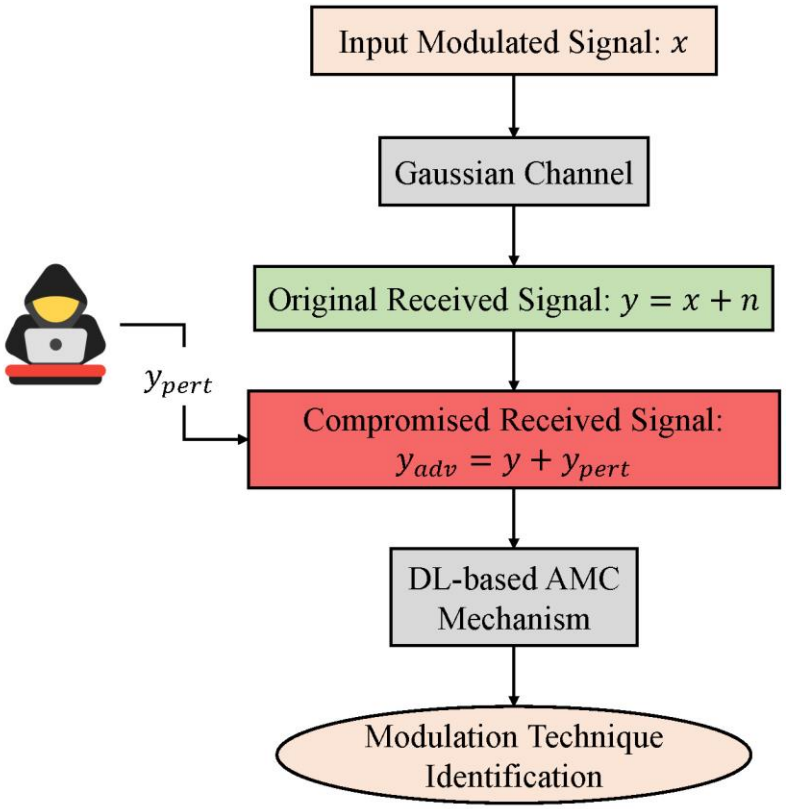
# PROBLEM FORMULATION




- We develop a practical and stealthy adversarial attack to compromise the automated decision-making process of a DL-powered AMC system.
  - ✓ We assume that the attacker is capable to launch an eavesdropping attack and obtain the information on the signal $y = x + n$ and the information on the DL model adopted by the AMC mechanism. Based on the information, the attacker calculates the perturbation $y_{pert}$ and launches a Man-In-The-Middle (MITM) attack to compromise the information at the receiver $y_{adv} = y + y_{pert}$.

Fig. 1: Overview of the AMC system considered in our work.

# PROBLEM FORMULATION




- We develop a practical and stealthy adversarial attack to compromise the automated decision-making process of a DL-powered AMC system.
  - ✓To formulate and implement our proposed adversarial attack, we leverage RADIOML 2018.01A dataset that was generated by using GNU radio.

Fig. 1: Overview of the AMC system considered in our work.

# AGENDA

- Introduction
- Problem formulation
- Proposed method
- Performance evaluations
- Conclusions

# PROPOSED METHOD

- We formulate the threat model of our proposed adversarial attack as follows:
  - ✓The attacker is assumed to launch eavesdrop attacks on the receiver to obtain: 1) the information of the received signal $y$ and 2) the information of the DL model used in the AMC mechanism.
  - ✓Based on the information obtained via eavesdropping attack, the attacker is able to implement the proposed adversarial attack algorithm to calculate the perturbation $y_{pert}$.
  - ✓The attacker is assumed to be able to launch a MITM attack on the receiver for compromising the signal $y$ with the perturbation $y_{pert}$.

# PROPOSED METHOD

- Our proposed adversarial attack is developed to achieve a practical tradeoff between attack impact and stealthiness. To achieve this goal, we formulate our proposed attack as:

$$\underset{y_{pert}}{\text{argmax}} \ \mathcal{D}\left(f\left(y + y_{pert}\right), f(y)\right)$$

$$\text{s.t. } \mathcal{P}\left(y_{pert}\right) \leq p$$

  - $\mathcal{D}(\cdot)$ is defined to measure the distance between the initial decision-making output of the AMC mechanism and the compromised decision-making output after launching the adversarial attack. In our work, it is formulated based on cross-entropy.
  - $f(\cdot)$ is the decision-making of the AMC mechanism for identifying the modulation technique.

# PROPOSED METHOD

- Our proposed adversarial attack is developed to achieve a practical tradeoff between attack impact and stealthiness. To achieve this goal, we formulate our proposed attack as:

$$\underset{y_{pert}}{\text{argmax}} \ \mathcal{D}\left(f\left(y + y_{pert}\right), f(y)\right)$$

$$\text{s.t. } \mathcal{P}\left(y_{pert}\right) \leq p$$

  - $\mathcal{P}\left(y_{pert}\right)$ is the perturbation power introduced when the attacker launches the MITM attack.
  - To ensure that the attack is stealthy and has a high success rate of bypassing the inherent attack detector in the receiver, the perturbation power $\mathcal{P}\left(y_{pert}\right)$ is constrained by a predetermined threshold $p$.

# PROPOSED METHOD

- In our current stage of research, the attacker is assumed to realize the proposed adversarial attack by employing FGSM, FGV, and BIM methods.
  - ✓FGSM: $y_{adv} = \left\{ y + g\left(\text{sign}\left(\nabla_y J(\theta, y, f(y))\right)\right) \middle| y \in Y \right\}$,
    - $\nabla_y J(\cdot)$ is the gradient of the loss with respect to the input feature $y$.
    - $g(\cdot)$ is defined to formulate the constraint in the proposed adversarial attack.

$$g(m) = \begin{cases} \sqrt{p}\,\frac{m}{\|m\|_2} & \text{if} \|m\|_2^2 > p \\ m & \text{otherwise} \end{cases}$$

  - ✓FGV: $y_{adv} = \left\{ y + g\left(\nabla_y J(\theta, y, f(y))\right) \middle| y \in Y \right\}$

# PROPOSED METHOD

- In our current stage of research, the attacker is assumed to realize the proposed adversarial attack by employing FGSM, FGV, and BIM methods.
  - ✓BIM: $y_{adv}^{N+1} = clip_{\epsilon}\left(y_{adv}^{N} + \text{sign}\left(\nabla_y J\left(\theta, y_{adv}^{N}, f(y)\right)\right)\right)$,
    - $y_{adv}^{0} = y, y \in Y$.
    - $clip_{\epsilon}(\cdot)$ is defined to formulate the constraint in the proposed adversarial attack.

$$clip_{\epsilon}(m) = \begin{cases} y + \sqrt{\frac{p}{\dim(y)}} & \text{if } m > y + \sqrt{\frac{p}{\dim(y)}} \\ y - \sqrt{\frac{p}{\dim(y)}} & \text{if } m < y - \sqrt{\frac{p}{\dim(y)}} \\ m & \text{otherwise} \end{cases}$$

# AGENDA

- Introduction
- Problem formulation
- Proposed method
- Performance evaluations
- Conclusions

# PERFORMANCE EVALUATIONS

- Since the work in this paper mainly focuses on developing and evaluating our proposed adversarial attack, for simplicity, we consider that the victim AMC mechanism has a satisfying performance when no adversarial attack is launched.
- By using the preprocessed data samples, we optimize the performances of the CNN-powered AMC mechanism and the GRU-based AMC mechanism.

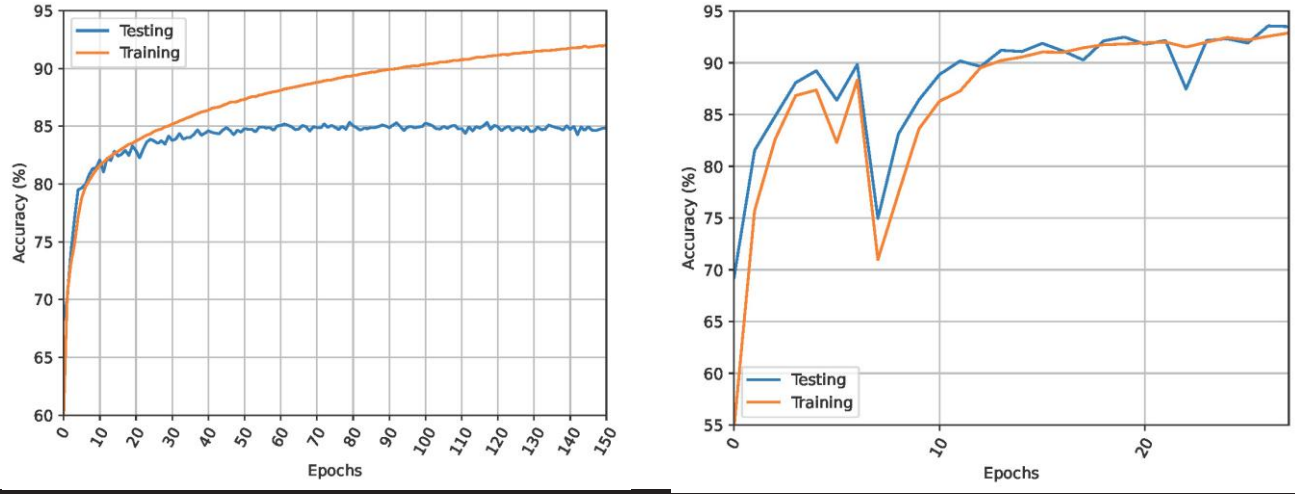



Fig. 2: Performances of the CNN-powered AMC mechanism and that of the GRU-powered AMC mechanism, respectively.

# PERFORMANCE EVALUATIONS

- By using our achieved CNN-powered AMC mechanism and GRU-powered AMC mechanism, we continue to evaluate the performance of our proposed practical and stealthy adversarial attack that is realized based on FGSM, FGV, and BIM.
  - ✓We consider the different values of the practical constraint $p$ of the perturbation power in the adversarial attack formulation from $-6\ dB$ to $3\ dB$.

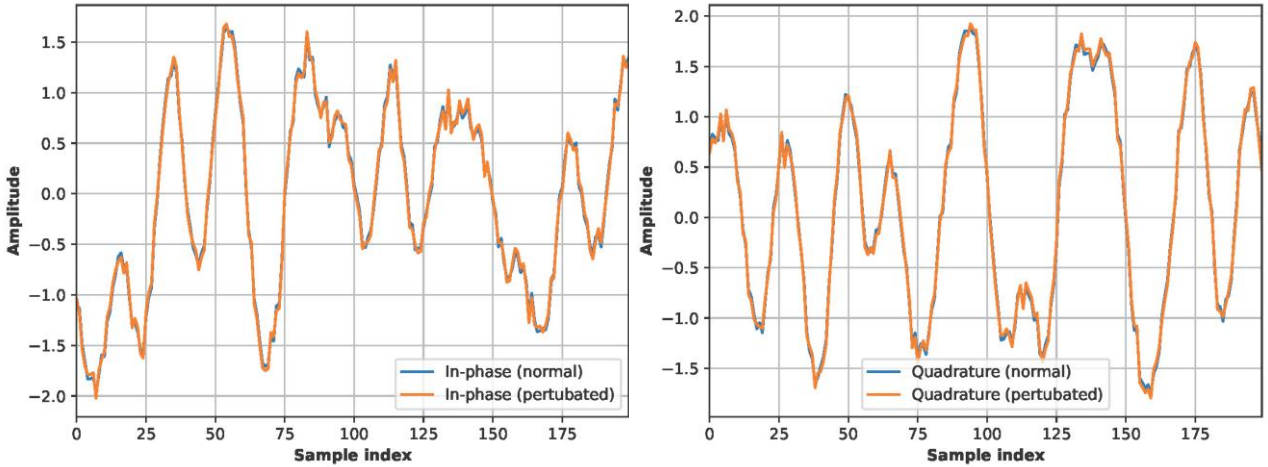



Fig. 3: Comparison of in-phase (left) and quadrature (right) components of a randomly selected data sample before and after launching our proposed BIM-based adversarial attack on the CNN-powered AMC mechanism ($p = 3dB$).

# PERFORMANCE EVALUATIONS

- By using our achieved CNN-powered AMC mechanism and GRU-powered AMC mechanism, we continue to evaluate the performance of our proposed practical and stealthy adversarial attack that is realized based on FGSM, FGV, and BIM.
  - ✓We consider the different values of the practical constraint $p$ of the perturbation power in the adversarial attack formulation from $-6\ dB$ to $3\ dB$.

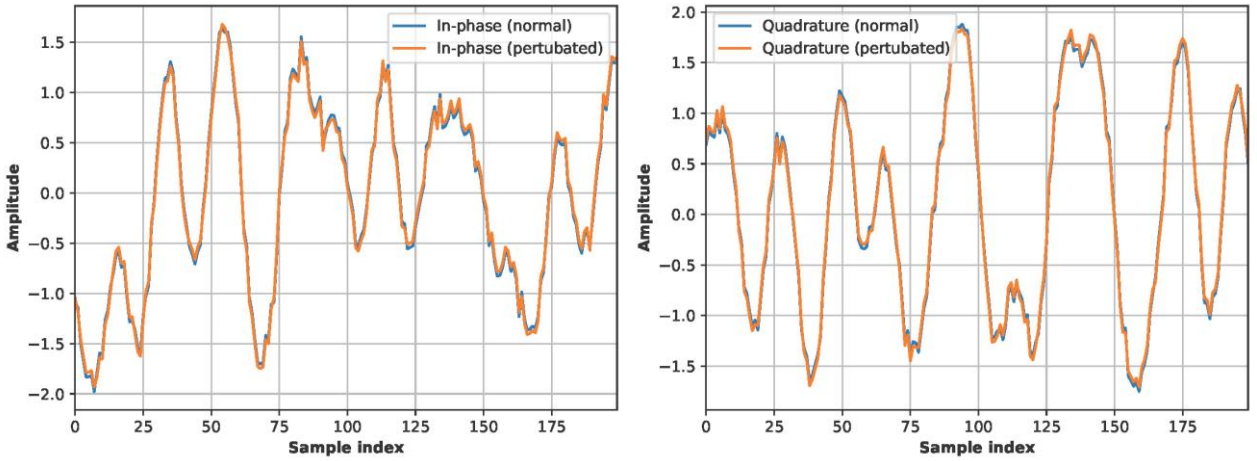



Fig. 4: Comparison of in-phase (left) and quadrature (right) components of a randomly selected data sample before and after launching our proposed BIM-based adversarial attack on the GRU-powered AMC mechanism ($p = 3dB$).

# PERFORMANCE EVALUATIONS

- The performances of our proposed adversarial attack on compromising the CNN-powered and GRU-powered AMC mechanism are presented.

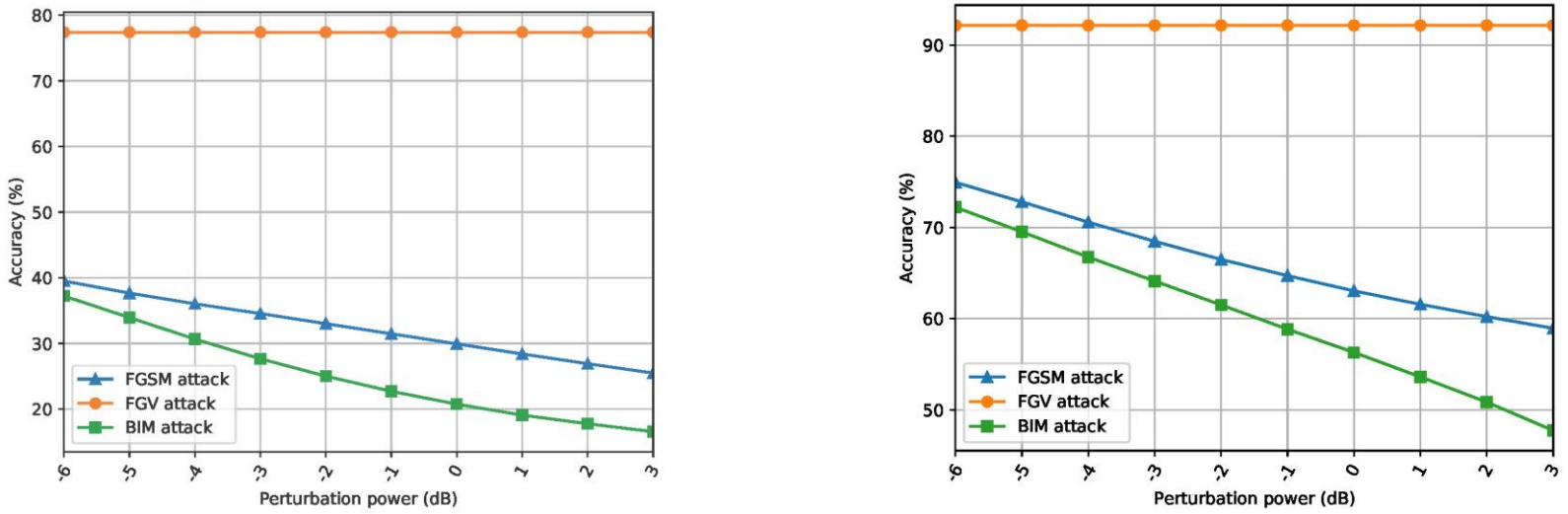
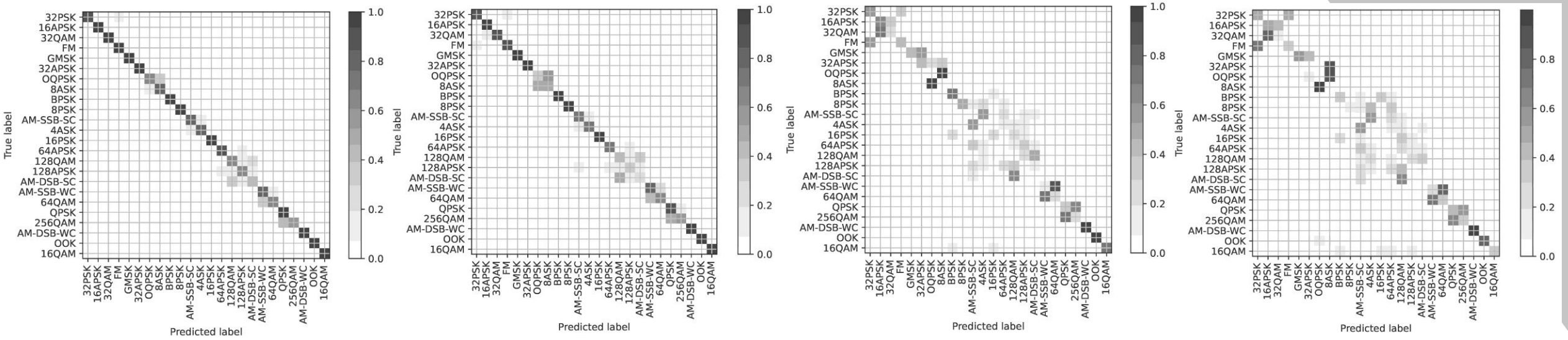



Fig. 5: Performances of our proposed adversarial attack on compromising the detection accuracy of the victim CNN-powered and GRU-powered AMC mechanisms, respectively: Left: Compromised accuracy of the victim CNN-powered AMC mechanism, and Right: Compromised accuracy of the victim GRU-powered AMC mechanism.

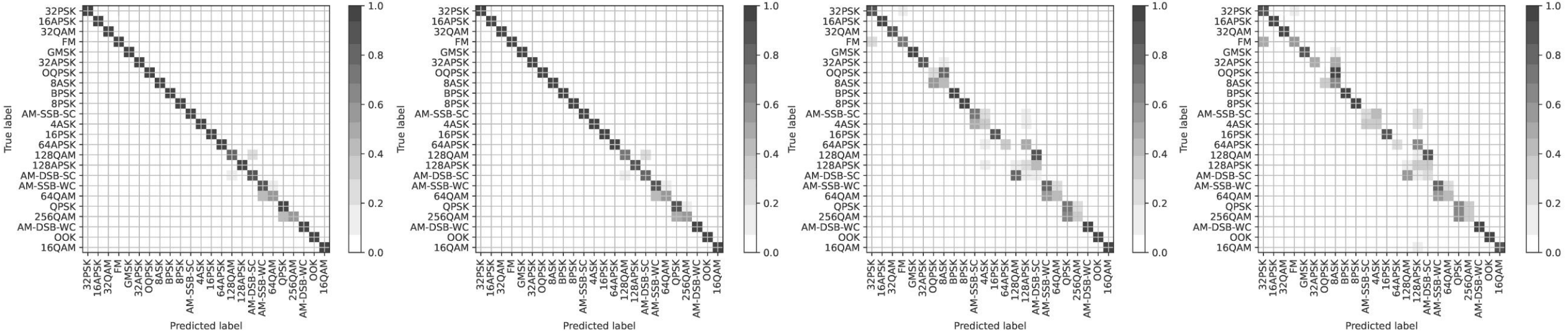
# PERFORMANCE EVALUATIONS




(a) (b) (c) (d)

Fig. 6: The compromised performances of the victim CNN-powered AMC mechanism by launching our proposed adversarial attacks with a practical constraint of the perturbation power $p = -2\ dB$: (a): No adversarial attack, (b): FGV-based adversarial attack, (c): FGSM-based adversarial attack, and (d): BIM-based adversarial attack.

# PERFORMANCE EVALUATIONS




(a) (b) (c) (d)

Fig. 7: The compromised performances of the victim GRU-powered AMC mechanism by launching our proposed adversarial attacks with a practical constraint of the perturbation power $p = -2\ dB$: (a): No adversarial attack, (b): FGV-based adversarial attack, (c): FGSM-based adversarial attack, and (d): BIM-based adversarial attack.

# AGENDA

- Introduction
- Problem formulation
- Proposed method
- Performance evaluations
- Conclusions

# CONCLUSIONS

- In this work, we focus on studying emerging adversarial attacks in the wireless communication domain. To mitigate this emerging attack, it can be beneficial to first adopt an adversarial mindset and formulate practical threat models of adversarial attacks in compromising wireless communication operations.
- In this paper, we focus on AMC operation in wireless communication and develop a practical and stealthy adversarial attack with three realizations based on FGSM, FGV, and BIM methods.
- The simulation results illustrate the effectiveness of our proposed adversarial attack model in achieving a practical tradeoff between accuracy reduction and stealthiness.

# THANK YOU!

# QUESTIONS?