# Adversarial Robustness Analysis of Deep Learning-Based Automatic Modulation Classification in Wireless Communication

Sunjun Hwang
Division of Software
Yonsei University
Wonju, Republic of Korea
sunjun7559012@yonsei.ac.kr

Eunho Choi
Division of Biomedical Engineering
Yonsei University
Wonju, Republic of Korea
agho@yonsei.ac.kr

Dohyun Hwang
Division of Software
Yonsei University
Wonju, Republic of Korea
ezwez1467@yonsei.ac.kr

***Abstract*—Deep learning-based Automatic Modulation Classification (AMC) has emerged as a critical technology for spectrum sensing and cognitive radio applications. However, the vulnerability of these models to adversarial attacks poses significant security concerns in wireless communication systems. This paper presents a comprehensive evaluation of adversarial robustness for a VTCNN2-based AMC model using the RadioML2016.10A dataset. We systematically analyze three representative adversarial attacks—FGSM, DeepFool, and C&W—and evaluate two defense mechanisms: adversarial training and Denoising Autoencoder (DAE). Our experimental results demonstrate that the baseline model is highly susceptible to adversarial perturbations, with accuracy dropping from 54.02% to as low as 10.80% under DeepFool attack. FGSM-based adversarial training improves robustness across multiple attacks, increasing accuracy under FGSM from 17.59% to 33.47% and improving robustness under DeepFool (10.80%→15.36%) and C&W (29.41%→40.67%), with a modest clean-accuracy drop (54.02→51.48%). In contrast, DAE preprocessing preserves clean accuracy(53.94%) but provides negligible improvement under FGSM(17.59→17.48) and only limited gains under iterative L2 attacks such as DeepFool (10.80%→21.16%) and C&W (29.41%→32.81%). These findings highlight the urgent need for robust defense mechanisms in deep learning-based wireless communication systems.**

***Keywords—automatic modulation classification, adversarial attack, deep learning, wireless security, robustness***

## I. INTRODUCTION

Automatic Modulation Classification (AMC) is a fundamental task in wireless communication systems, enabling receivers to automatically identify the modulation scheme of incoming signals without prior knowledge. This capability is essential for spectrum sensing in cognitive radio networks, electronic warfare, and signal intelligence applications. With the advent of deep learning, convolutional neural networks (CNNs) have demonstrated remarkable performance in AMC tasks, achieving high classification accuracy across various signal-to-noise ratio (SNR) conditions.

However, deep neural networks are known to be vulnerable to adversarial examples—carefully crafted input perturbations that are imperceptible to humans but can cause misclassification. In the context of wireless communications, adversarial attacks on AMC systems could enable malicious actors to evade detection, disrupt spectrum sensing, or compromise communication security. Despite the critical implications, the adversarial robustness of deep learning-based AMC models remains underexplored.

In this paper, we conduct a systematic evaluation of adversarial vulnerabilities in a VTCNN2-based AMC model. Our contributions are threefold: (1) We evaluate three representative adversarial attacks—Fast Gradient Sign Method (FGSM), DeepFool, and Carlini & Wagner (C&W)—on the RadioML2016.10A benchmark dataset; (2) We analyze the effectiveness of adversarial training as a defense mechanism; (3) We investigate the limitations of Denoising Autoencoder (DAE) based defense for wireless signal protection.

## II. RELATED WORK

### A. Deep Learning for AMC

O'Shea et al. pioneered the application of deep learning to AMC with their VTCNN2 architecture, demonstrating that CNNs can effectively learn discriminative features from raw In-phase/Quadrature (IQ) signal data [1]. Subsequent works have explored various architectures including residual networks, recurrent neural networks, and attention mechanisms to further improve classification performance.

### B. Adversarial Attacks on Wireless System

Recent studies have begun investigating adversarial vulnerabilities in wireless deep learning systems. Sadeghi and Larsson demonstrated that adversarial perturbations can significantly degrade AMC performance [2]. Flowers et al. explored physical-layer adversarial attacks in over-the-air scenarios [3]. However, comprehensive evaluations

comparing multiple attack methods and defense strategies remain limited.

## III. METHODOLOGY

### A. Dataset

We utilize the RadioML2016.10A dataset, a widely adopted benchmark for AMC research [4]. The dataset contains over 220,000 samples of IQ signal data with shape (2, 128), representing 11 modulation types across 20 SNR levels ranging from −20 dB to +18 dB in 2 dB increments. Table I summarizes the dataset characteristics.

TABLE I. RADIOML.2016.10A DATASET OVERVIEW

| Parameter | Value |
|---|---|
| Signal Format | IQ, Shape (2, 128) |
| Total Samples | 220,000+ |
| Modulation Types | 11 |
| SNR Range | -20 to +18dB |
| Train / Test | 154,000 / 66,000 |

### B. Baseline Model Architecture

We employ a modified VTCNN2 architecture as our baseline classifier. The model takes normalized IQ signals reshaped to (1, 2, 128) as input. The architecture consists of two convolutional blocks: Conv2d(1→256) and Conv2d(256→80), each followed by batch normalization, ReLU activation, and max pooling. Dropout (0.5) is applied

before fully connected layers (2560 → 256 → 11). The model contains approximately 500K parameters.

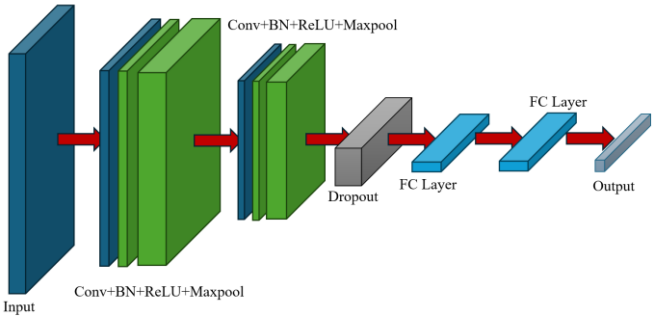


Fig. 1. VTCNN2 model architecture: Input(1, 2, 128) → Conv2d+BN+ReLU+Maxpool → Conv2d+BN+ReLU+Maxpool → Dropout → FC layers → Output(11).

Implementation details: The baseline VTCNN2 model is trained using the Adam optimizer with a learning rate of $5 \times 10^{-4}$ and a batch size of 256 for 20 epochs. The dataset is split into 154,000 training samples and 66,000 test samples, following the standard RadioML2016.10A protocol. Input IQ samples are normalized to the range [0, 1] before being fed into the network.

### C. Adversarial Attack Methods

- **FGSM[5]**: Single-step adversarial attack that perturbs the input in the direction of the sign of the gradient of the loss with respect to the input. It is computationally efficient and widely used as a baseline gradient-based attack.

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y)) \quad (1)$$

- **DeepFool[6]**: an iterative attack that approximates the classifier as locally linear and moves the input toward the closest decision boundary. At each iteration, it computes the minimal perturbation required to cross the boundary, resulting in a perturbation with minimal $L_2 - norm$.

$$x_{i+1} = x_i + \left\{\frac{|f(x_i)|}{||\nabla_x(f(x_i)||_2^2}\right\}\nabla_x f(x_i) \quad (2)$$

- **C&W[7]**: The Carlini–Wagner attack formulates adversarial example generation as an optimization problem. It searches for the smallest perturbation that causes misclassification while ensuring the perturbed sample satisfies classifier constraints. It is considered one of the strongest gradient-based white-box attacks.

$$\min_{\delta} ||\delta||_2^2 + c \cdot f(x+\delta) \quad (3)$$

$$f(x') = \max\left(\max_{i \neq t}(Z(x')_i) - Z(x')_t, -\kappa\right) \quad (4)$$

- **Attack configuration.** FGSM uses an $L_\infty$ perturbation budget $\epsilon = 0.005$ on normalized [0,1] IQ inputs. DeepFool is configured with 50 steps and overshoot 0.02. The C&W($L_2$) attack uses $c = 1e-4, \kappa = 0,30$ steps, and learning rate 0.005.

### D. Defense Mechanisms

- **Adversarial Training [8]**: Training data is augmented with FGSM-generated adversarial examples (ε = 0.005). The model is trained on mixed clean and adversarial samples for 5 epochs. We use the Adam optimizer with learning rate 5e-4 for 5 epochs, training on a mixture of clean and FGSM samples ($\epsilon = 0.005$).

- **Denoising Autoencoder (DAE)**: A lightweight convolutional autoencoder trained to reconstruct clean signals from adversarial inputs as a preprocessing defense.

## IV. EXPERIMENTAL RESULTS

### A. Baseline Performance

The baseline model achieves an overall accuracy of 54.02% across all SNR levels. Table II shows accuracy by SNR. Performance improves significantly with higher SNR, reaching 81.6% at 10 dB. At low SNR (−20 to −12 dB), accuracy is near random (~9–12%). For SNR ≥ 0 dB, average accuracy is approximately 80%.

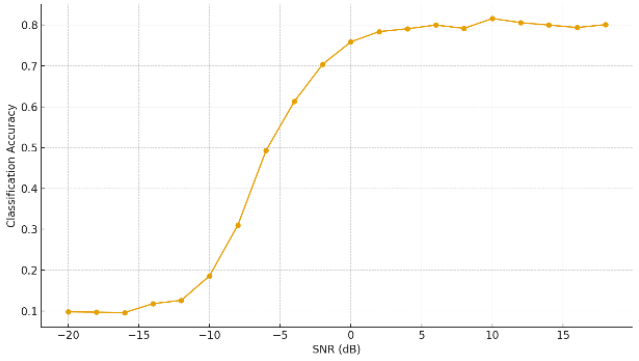


Fig. 2. Classification accuracy versus SNR. Sigmoidal curve from ~0.09 at −20 dB to ~0.80 plateau for SNR ≥ 0 dB

TABLE II. BASELINE ACCURACY BY SNR

| SNR | ACC. | SNR | ACC. |
|---|---|---|---|
| -10 | 0.186 | 4 | 0.791 |
| -8 | 0.310 | 6 | 0.800 |
| -6 | 0.493 | 8 | 0.792 |
| -4 | 0.613 | 10 | 0.816 |
| -2 | 0.704 | 12-18 | 0.79-0.81 |
| 0 | 0.759 | **Overall** | 0.5402 |
| 2 | 0.784 | Avg(SNR≥0) | 0.80 |

### *B. Adversarial Attack Evaluation*

Table III reports accuracy under attacks. The baseline drops from 0.5402 (clean) to 0.1759 (FGSM, ε = 0.005), 0.1080(DeepFool), and 0.2941 (C&W), where DeepFool is the most effective.

### *C. Defense Mechanism Evaluation*

FGSM-based adversarial training improves robustness not only to FGSM (0.1759 → 0.3347) but also to DeepFool (0.1080→0.1536) and C&W (0.2941→0.4067), with a modest clean-accuracy drop (0.5401 → 0.5148). DAE preprocessing preserves clean accuracy (0.5394) but yields negligible improvement under FGSM (0.1759→0.1748) and limited gains under iterative $L_2$ attacks (0.1080 → 0.2116 for DeepFool, 0.2941→0.3281 for C&W).

## V. DISCUSSION

Our results reveal critical insights about adversarial robustness in AMC systems. The substantial accuracy degradation (54% → 10 – 29%) confirms high vulnerability to adversarial perturbations, posing serious security concerns for cognitive radio applications.

DeepFool's superior effectiveness stems from its iterative optimization that precisely identifies minimal perturbations to cross decision boundaries. In IQ signal space, learned features are sensitive to targeted perturbations in specific input directions.

Adversarial training's effectiveness (83% relative improvement) demonstrates that training-time defenses provide meaningful protection. However, the clean accuracy decrease highlights the fundamental accuracy-robustness trade-off requiring careful balance in deployments. Robustness is strongly SNR-dependent, for SNR ≥ 0 dB, the baseline accuracy decreases from approximately 0.79 in clean signals to about 0.31 under FGSM, whereas adversarial training restores it to around 0.51. DeepFool remains the strongest attack across SNRs, and adversarial training only partially mitigates its impact.

The DAE failure is noteworthy. Unlike image domains, wireless IQ signals exhibit high sensitivity to small amplitude and phase perturbations, and discriminative modulation patterns are often embedded in fine-grained signal structures. Reconstruction statistics indicate that the DAE behaves almost as an identity mapping on clean inputs, yielding very small reconstruction error, while simultaneously compressing the input dynamic range. This MSE-driven smoothing effect can attenuate modulation-specific features together with adversarial perturbations. As a result, DAE preprocessing preserves clean accuracy but provides negligible robustness improvement under FGSM and only limited gains under iterative L2 attacks such as DeepFool and C&W. These observations suggest that reconstruction-based defenses optimized for pixel-wise fidelity are insufficient for wireless signal protection, and that effective defenses should explicitly account for the structural characteristics of modulation patterns.

## VI. CONCLUSION

This paper presents a comprehensive adversarial robustness evaluation for deep learning-based AMC. Experiments on RadioML2016.10A demonstrate VTCNN2 models are highly susceptible to attacks, with DeepFool achieving the most severe degradation (10.80%). FGSM-based adversarial training provides effective defense across multiple attacks, while DAE preserves clean accuracy but offers limited robustness improvements.

These findings underscore the critical need for robust deep learning models in wireless systems. Future directions include certified defenses, adversarial transferability analysis, and wireless-specific defense strategies.

TABLE III. CLASSIFICATION ACCURACY UNDER ATTACKS AND DEFENSES (FGSM: $\epsilon = 0.005$)

| Model/Defense | Clean | FGSM | DeepFool | C&W |
|---|---|---|---|---|
| Baseline | 0.5402 | 0.1759 | 0.1080 | 0.2941 |
| Adv. Training (FGSM) | 0.5148 | 0.3347 | 0.1536 | 0.4067 |
| DAE (preproc.) | 0.5394 | 0.1748 | 0.2116 | 0.3281 |

Note that $\epsilon$ applies only to FGSM; DeepFool and C&W follow the configurations described in Section III.C

## REFERENCES


[1] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in Proc. Int. Conf. Eng. Appl. Neural Netw., 2016, pp. 213-226.

[2] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," IEEE Wireless Commun. Lett., vol. 8, no. 1, pp. 213-216, 2019.

[3] B. Flowers, R. M. Buehrer and W. C. Headley, "Evaluating Adversarial Evasion Attacks in the Context of Wireless Communications," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1102-1113, 2020.

[4] T. J. O'Shea and N. West, "Radio machine learning dataset generation with GNU radio," in Proc. GNU Radio Conf., vol. 1, no. 1, 2016.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent., 2015.

[6] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2574-2582.

[7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Secur. Privacy, 2017, pp. 39-57.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. Learn. Represent., 2018.