Towards Intelligent Security Mechanisms for Connected Things

Paulo Freitas de Araujo-Filho[1,2], **Divanilson R. Campelo**[1], **Georges Kaddoum**[2]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil

[2]Electrical Engineering Department
École de Technologie Supérieure (ÉTS), Université du Québec
Montreal, QC, Canada

{pfreitas,dcampelo}@cin.ufpe.br, georges.kaddoum@etsmtl.ca

Abstract. *The widespread adoption of connected devices and the adoption of machine learning enable attackers to launch several cyber-attacks and adversarial attacks. Therefore, the goals of this thesis are to investigate and develop cutting-edge solutions to enhance the security of systems by effectively and efficiently detecting cyber-attacks while also defending systems that rely on ML from adversarial attacks. The main results of our thesis comprehend multiple awards, the publication of eight papers in prestigious journals, three conference papers, two patents, and one software registration. Furthermore, our research has been recognized and awarded as one of the two 2022 Microsoft Research Ph.D. Fellowship recipients in Security, Privacy, and Cryptography worldwide.*

Resumo. *A ampla adoção de dispositivos conectados e de modelos de aprendizagem de máquina permite que atacantes realizem diversos ciberataques e ataques adversariais. Assim, os objetivos desta tese são investigar e desenvolver soluções de ponta para aprimorar a segurança de sistemas, detectando de maneira eficaz e eficiente ciberataques e defendendo-os de ataques adversariais. Os seus principais resultados representam múltiplos prêmios, a publicação de oito artigos em revistas de prestígio, três artigos em conferências, duas patentes e um registro de software. Além disso, nossa pesquisa foi premiada como um dos dois únicos ganhadores em todo o mundo do Microsoft Research Ph.D. Fellowship em 2022 na área de Segurança, Privacidade e Criptografia.*

1. Introduction and Motivation

The increasing growth of connected devices, which comprehend the Internet-of-things (IoT), is changing the way we interact with our surroundings. This connected environment is expected to even further increase with the deployment of the fifth-generation (5G) and beyond mobile networks. On the other hand, the broadcast nature of wireless communications enables attackers to eavesdrop and inject malicious data into the network and launch several cyber-attacks [Pourranjbar et al. 2022]. Moreover, while machine learning (ML) is being largely adopted in many applications, it also introduce new risks and vulnerabilities. Adversarial attacks craft and introduce small perturbations that fool ML models into making wrong decisions, which then may significantly impact the security of systems and

networks just as cyber-attacks do [Yuan et al. 2019]. Therefore, despite numerous security solutions available, the IoT's physical constraints, highly heterogeneous environment, and the use of ML impose new security challenges [Pourranjbar et al. 2023].

1.1. Problem Statement

Since new cyber-attacks are constantly launched and obtaining labeled attack data is very challenging, intrusion detection systems (IDSs) need to rely on unsupervised learning techniques to detect both known and unknown attacks and to not require labeled data. However, most existing unsupervised IDSs cannot deal with correlations in multivariate time series that are extensively present in IoT data, increasing their false positive rates [Nisioti et al. 2018]. Therefore, it is necessary to propose novel unsupervised IDSs that simultaneously achieve low false positive and negative rates.

Moreover, since cyber-attacks need to be stopped before causing damage, the detection time of IDSs needs to be as short as possible. However, most state-of-the-art IDSs have long detection times for relying on long short-term memory (LSTM) neural networks. Although LSTM networks improve detection by considering time dependencies among data, their limited capacity to parallelize computations increases detection time [Pourranjbar et al. 2023]. Therefore, it is necessary to investigate other architectures that consider time dependencies among data while allowing the fast detection of attacks.

Finally, ML has been shown vulnerable to adversarial attacks, which can cause severe security issues to systems that rely on them. Adversaries can, for example, force ML-based modulation classifiers used in wireless communications to produce incorrect outputs and interrupt communication. However, only a few works have proposed techniques to defend connected objects from such attacks, most of which only marginally reduce the impact of the attacks [Yuan et al. 2019]. Therefore, further investigations are necessary to ensure the security of systems against adversarial attacks.

1.2. Related Works

After conducting an extensive literature review, we verified that although many security solutions exist, there are still several issues to be tackled. While our complete literature review can be found in our thesis and papers, due to page limitations, we summarize here the concluding remarks and open challenges that we identified and aim to solve with our thesis:

- While IDSs should not rely on labelled data, most of them present high false positive rates and struggle with the time required to detect intrusions. Thus, it is necessary to propose new detection solutions that reduce the detection time and achieve low false positive and false negative rates.
- While LSTM networks are heavily used by state-of-the-art IDSs, they present several drawbacks that put in doubt their status as the standard architecture for sequence modeling tasks. Thus, it is necessary to investigate novel strategies for considering time-dependencies among data.
- Although adversarial attacks may significantly compromise the security of systems that rely on ML, their study is still in its early stages. Thus, it is necessary to investigate the impact of adversarial attacks on different application domains and propose techniques to enhance systems' security against them.

1.3. Objectives

Although cyber-attacks and adversarial attacks represent different techniques for compromising security, their effects are the same, as they can severely compromise security. Hence, given their potential impact, the hypothesis that guides our research is whether artificial intelligence enhances security by effectively and efficiently detecting attacks or harms security due to the vulnerabilities it adds. Therefore, in our research, we aim to advance the state-of-the-art in the security field by addressing the aforementioned identified challenges. Our main goal is to enhance the security of systems by effectively and efficiently detecting cyber-attacks while also defending systems that rely on ML from adversarial attacks. To achieve our goal, we define the following four specific objectives:

1. Propose an unsupervised IDS that reduces the detection time of the current state-of-the-art solutions, making it more suitable for latency-constrained applications.
2. Propose an unsupervised IDS that considers time-dependencies among data without relying on LSTM networks, such that their drawbacks are avoided.
3. Propose an adversarial attack technique and investigate the extent to which it may jeopardize security by compromising the availability of systems.
4. Investigate and propose a defense technique that protects ML-based systems from adversarial attacks.

1.4. Contributions

In this thesis, we advance the state-of-the-art in the security field by considering the cyber-attacks and adversarial attacks problems. Our contributions include:

- Four prestigious grants and awards;
- Eight papers published in prestigious journals;
- Three conference papers;
- Two patents;
- One software registration;
- Three publications for the general public.

While the complete references and additional information of our research accomplishments are presented in the attached subproducts document, we highlight that our research has been recognized and awarded as one of the two 2022 Microsoft Research Ph.D. Fellowship recipients in Security, Privacy, and Cryptography, from a total of 36 recipients in all areas worldwide.

Furthermore, we highlight that our main contributions can be divided in two parts. The first part concerns the use of ML for intrusion detection with the proposal of unsupervised IDSs to accomplish our first two specific objectives. It comprehends the contributions in [J4] and [J2], patents [P1] and [P2], and Chapters 3 and 4 of our thesis. The second part concerns the security of systems that use ML, and comprehends our last two specific objectives. It comprehends the contributions in [J3] and [J1], and Chapters 5 and 6 of our thesis.

2. ML-Based Intrusion Detection

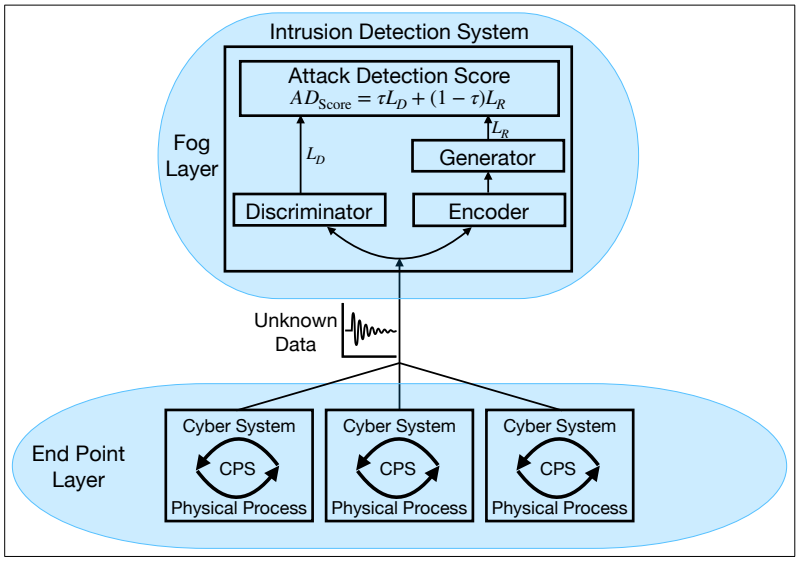
2.1. Intrusion Detection for Cyber-Physical Systems using Generative Adversarial Networks in Fog Environment

In Chapter 3, we propose a novel unsupervised IDS that detects known and unknown attacks using LSTM networks to consider time dependencies among data and generative

adversarial networks (GANs). While the GAN discriminator directly evaluates whether a sample is an intrusion, the generator is used to compute a reconstruction loss that can be combined with the discriminator's output to improve detection rates. Moreover, we propose an Encoder neural network that accelerates the reconstruction loss computation and significantly reduces the detection latency by eliminating the need for solving an optimization problem during the detection of intrusions. Furthermore, to reduce even more the detection latency, our IDS takes advantage of the fog-computing paradigm, being deployed in the fog as a virtual function. Figure 1 exhibits the architecture of the proposed IDS detection model. In a nutshell, the main contributions of the work in this chapter are:

1. An unsupervised anomaly-based IDS using GAN, which is capable of detecting unknown attacks and overcomes the challenge of obtaining labels.
2. Evaluation of the individual contribution of the GAN discrimination and reconstruction losses in the detection of cyber-attacks to improve the detection rates.
3. Proposal of a novel and faster method for inverting the GAN generator, which is useful for latency constrained classification and retrieval tasks.
4. Proposal of a fog-based architecture for our IDS, which enables our security solution to take advantage of the low-latency of fog nodes-based applications.

Figure 1. Proposed FID-GAN Detection Model



2.1.1. Methodology, datasets, and results

In our experiments, we evaluated both the detection rates and detection latency of our IDS when relying on only the discrimination loss, on only the reconstruction loss, and on a combination of the discrimination and reconstruction losses. We considered three datasets: the SWaT and the WADI for sensor measurements of a water treatment plant, and the NSL-KDD for network traffic. Our experiments show that our proposed IDS achieves detection rates that are higher than those of two other state-of-the-art IDSs, namely [Li et al. 2019] and [Zenati et al. 2018], while also being at least 5.5 times faster than the IDS proposed in [Li et al. 2019] when considering only the reconstruction loss. Therefore, we verified that GANs have an important role as an unsupervised technique

for detecting attacks and that our proposed solution is much more suitable for latency constrained applications, such as the detection of cyber-attacks. Please refer to Chapter 3 of our thesis or to the paper [J4] for the complete results of our solution.

2.1.2. Associated subproducts

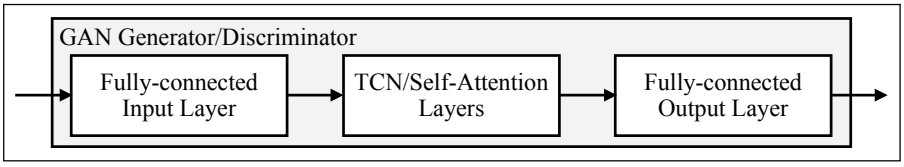
The contents of this chapter resulted in the subproduct [J4], as listed in our subproducts document. It represents a paper published in the prestigious IEEE Internet of Things Journal with an impact factor of 10.6. Additionally, this work has been cited 106 times so far.

2.2. Unsupervised GAN-Based Intrusion Detection System Using Temporal Convolutional Networks and Self-Attention

Since many attacks have multiple steps and are launched from different applications and devices, Chapter 4 concerns different strategies for considering time dependencies among data in the detection of attacks. In contrast to most state-of-the-art IDSs, we propose a novel unsupervised GAN-based IDS that uses temporal convolutional networks (TCNs) and self-attention as a replacement for LSTM networks. TCNs and self-attention enable more computation parallelization, have a constant number of sequentially executed operations, and have been shown to yield more accurate results than LSTM networks in specific sequence modeling tasks. Figure 2 shows the proposed high-level architectures of the GAN generator and discriminator. Moreover, we conduct a comparative evaluation of different TCN and self-attention GAN architectures so that different trade-offs between detection rates and detection times are achieved. In summary, the main contributions of our proposed TCN/self-attention GAN-based IDS are:

1. An unsupervised GAN-based IDS that is capable of detecting both known and zero-day attacks without relying on labeled attack data, which is difficult and sometimes impossible to obtain.
2. Experiments using TCNs and self-attention in a GAN to detect cyber-attacks with better detection results than existing GAN-based IDSs.
3. An evaluation of the trade-off between detection rates and detection times for different TCN and self-attention GAN architectures so that our proposed IDS can be configured to satisfy different requirements.

Figure 2. The GAN generator and discriminator architectures



2.2.1. Methodology, datasets, and results

Since in this work our main concern is to detect multi-step distributed attacks, such as modern distributed denial-of-service (DDoS) attacks, we use the publicly available CICDDoS2019 dataset to evaluate our proposed IDS. This dataset, which is provided by

the Canadian Institute for Cybersecurity (CIC) and the University of New Brunswick (UNB), contains benign traffic data and several of the most modern and common DDoS attacks types. In our experiments, we verify that our proposed approach successfully replaces LSTM networks for attack detection and achieves better detection results, surpassing the results of two state-of-the-art GAN-based IDSs: [Zenati et al. 2018] and [Freitas de Araujo-Filho et al. 2021]. Moreover, we verify the trade-off between detection rates and detection times for different configurations of our IDS so that our solution can be configured to satisfy different requirements depending on whether it is more important to achieve higher detection rates or shorter detection times. Precisely, our IDS achieves the highest detection rates and the longest detection times when using self-attention, and the lowest detection rates and the shortest detection times when using a single TCN block. Please refer to Chapter 4 of our thesis or to the paper [J2] for the complete results of our solution.

2.2.2. Associated subproducts

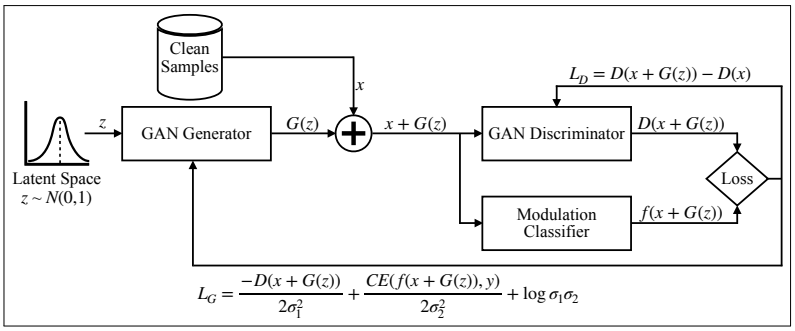
The contents of this chapter resulted in the subproduct [J2], as listed in our subproducts document. It represents a paper published in the prestigious IEEE Transactions on Network and Service Management with an impact factor of 5.3. Additionally, this work has been cited 11 times so far. Furthermore, this work has also led to the patents [P1] and [P2], also listed in our subproducts document.

3. Security for ML-Based Systems

3.1. Multi-Objective GAN-Based Adversarial Attack Technique for Modulation Classifiers

In Chapter 5, we verify that the existing adversarial attack techniques either require complete knowledge about the victim classifier's model, which is an unrealistic assumption, or take too long to craft adversarial perturbations. Hence, we propose a novel adversarial attack technique, which overcomes such limitations, for assessing the risks of using ML-based modulation classifiers in wireless communications and contributing for the development of classifiers that are robust against adversarial attacks. The main contributions of our work are as follows: First, we combine GANs and multi-task loss to generate adversarial samples, by simultaneously optimizing their ability to cause wrong classifications and not being perceived. While we modify the GAN original formulation so that it produces adversarial perturbations rather than samples similar to those of a training set, the multi-task loss allow us to simultaneously optimize multiple conditions and achieve both the success of the adversarial attack and its imperceptibility. Second, we reduce the accuracy of modulation classifiers more and craft adversarial samples in a shorter time than existing techniques while following the decision-based black-box scenario. Third, we propose an input-agnostic adversarial attack technique that does not depend on the original samples to craft perturbations. It allows adversarial perturbations to be prepared in advance, further reducing the time for executing the adversarial attack. Finally, our work verifies that modulation classifiers are at an increased risk and urgently need to be enhanced against adversarial attacks. Figure 3 shows the training model of our proposed adversarial attack technique.

Figure 3. Our proposed training model



3.1.1. Methodology, datasets, and results

We use the RADIOML 2016.10A dataset and VT-CNN2 modulation classifier designed by DeepSiG and publicly available in [O'Shea et al. 2016, O'Shea and West 2016] to evaluate our proposed adversarial attack technique. The dataset is constructed by modulating and exposing signals to an additive white Gaussian noise (AWGN) channel that includes sampling rate offset, random process of center frequency offset, multipath, and fading effects, as described in [O'Shea et al. 2016, O'Shea and West 2016]. After modulation and channel modeling, the signals are normalized and packaged into 220,000 samples of in-phase and quadrature components with length 128, each associated with a modulation scheme and a signal-to-noise ration (SNR). The VT-CNN2 modulation classifier, on the other hand, relies on deep convolutional neural networks and classifies samples among the eleven modulation schemes in the dataset. It works as the classifier that will be the victim of our attack and of two other baseline adversarial attacks with which we compare our results: [Moosavi-Dezfooli et al. 2017] and [Sadeghi and Larsson 2019].

Our experiments show that it is possible to quickly craft small imperceptible perturbations that severely compromise modulation classifiers' accuracy and hence wireless receivers' performance. Precisely, our proposed attack technique reduces the accuracy of the VT-CNN2 modulation classifier more than a jamming attack and two other state-of-the-art adversarial attack techniques, namely [Moosavi-Dezfooli et al. 2017] and [Sadeghi and Larsson 2019], while generating adversarial samples at least 335 times faster than them. Therefore, it is urgently necessary to enhance deep learning-based modulation classifiers' robustness against adversarial attacks. Please refer to Chapter 5 of our thesis or to the paper [J3] for the complete results of our proposed technique.

3.1.2. Associated subproducts

The contents of this chapter resulted in the subproduct [J3], as listed in our subproducts document. It represents a paper published in the prestigious IEEE Communications Letters with an impact factor of 4.1. Additionally, this work has been cited 13 times so far.

3.2. Defending Wireless Receivers Against Adversarial Attacks on Modulation Classifiers

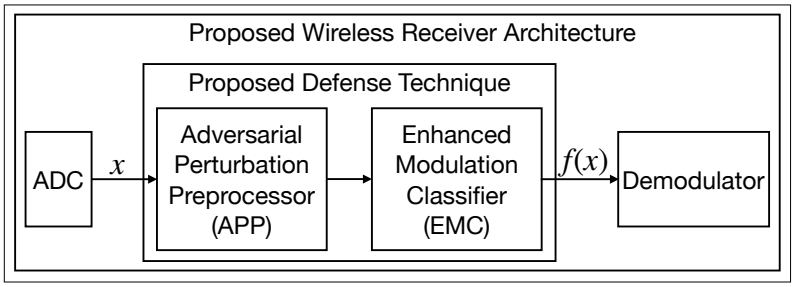
Given the risks and damage that adversarial attacks may cause, in Chapter 6 we propose a defense technique for protecting modulation classifiers from adversarial attacks so that

those attacks do not harm the availability of wireless communications. Our proposed defense technique is threefold. First, the amount of adversarial perturbation is estimated by relying on a denoising autoencoder (DAE) that has been specially trained to remove Gaussian noise and adversarial perturbations. Then, signals with considerable perturbations are preprocessed using the DAE to remove those undesirable attributes. Signals with small amounts of noise and adversarial perturbations, on the other hand, are not preprocessed as the DAE could introduce errors that are more significant than the perturbations. Finally, the signal's modulation scheme is identified with an enhanced modulation classifier (EMC) that has been trained using noisy and adversarial samples to make it resistant to sample variation. Figure 4 shows our proposed architecture. The analog-to-digital converter (ADC) forwards the received samples to a proposed adversarial perturbation preprocessor (APP) module, which contains our proposed DAE. Then, it processes the received samples and forwards them to the EMC module, which classifies the samples and indicates the recognized modulation scheme to the receiver's demodulator.

Compared to existing defense schemes, our proposed solution's first major technical improvement is our technique for estimating and removing adversarial perturbations, which significantly alleviates the burden on the classifier. Our proposed solution's second major technical improvement is its ability to enhance modulation classifiers' resistance to adversarial attacks while requiring only adversarial samples crafted using a single fast attack technique that is able to generalize other techniques. In a nutshell, the main contributions of our work are as follows:

1. We propose a DAE that has been specially trained to estimate and remove noise and adversarial perturbations from modulated signals.
2. We propose an enhanced modulation classifier (EMC) that is resistant to a variety of adversarial attack techniques.
3. We propose a novel wireless receiver architecture that is resistant to adversarial attacks by combining our proposed DAE and EMC to remove adversarial perturbations and make the classifier less affected by them.

Figure 4. Proposed wireless receiver architecture



3.2.1. Methodology, datasets, and results

To evaluate our proposed defense technique, we relied on the same dataset and modulation classifier used in Chapter 5, i.e., the RADIOML 2016.10A dataset and the VT-CNN2 modulation classifier [O'Shea et al. 2016, O'Shea and West 2016]. On the other hand, since we now propose a defense technique, we considered the worst-case scenario

of three white-box adversarial attacks, namely fast gradient sign method (FGSM), projected gradient descent (PGD), and momentum iterative method (MIM). That is, we now evaluate whether our proposed defense technique can protect the VT-CNN2 modulation classifier against the FGSM, PGD, and MIM attacks, which have complete knowledge about the victim's classifier, and compare our results to those of two state-of-the-art defense techniques: [Zhang et al. 2022] and [Zhang et al. 2021]. Our experiments show that our proposed technique significantly diminishes the accuracy reduction caused by adversarial attacks on modulation classifiers, and outperforms [Zhang et al. 2022] and [Zhang et al. 2021] by at least 18 percentage points. Therefore, we verified that better defense results are achieved by simultaneously removing adversarial perturbations and making classifiers less sensitive to them. Please refer to Chapter 6 of our thesis or to the paper [J1] for the complete results of our solution.

3.2.2. Associated subproducts

The contents of this chapter resulted in the subproduct [J1], as listed in our subproducts document. It represents a paper published in the prestigious IEEE Internet of Things Journal with an impact factor of 10.6.

4. Conclusion and Thesis Impact

As the increasing number of connected devices and the use of ML introduce new security challenges, our thesis proposed new strategies and techniques to protect connected things against cyber-attacks and adversarial attacks. We focused on developing novel IDSs that effectively and efficiently detect cyber-attacks, and defense techniques to mitigate the impacts of adversarial attacks. Since cyber-attacks and adversarial attacks may compromise the reliability of systems and jeopardize people's safety, our research outcomes are expected to significantly contribute to securing systems and networks, benefiting people, industries, and governments. In that sense, our research has been recognized and awarded the Microsoft Research Ph.D. Fellowship, being one of the two selected researches in Security, Privacy, and Cryptography worldwide in 2022, as well as by the Fonds de recherche du Québec. Furthermore, our thesis results are featured in top-tier venues in the field, as well as in two international patents published under the Patent Cooperation Treaty.

References

Freitas de Araujo-Filho, P., Kaddoum, G., Campelo, D. R., Gondim Santos, A., Macêdo, D., and Zanchettin, C. (2021). Intrusion Detection for Cyber–Physical Systems Using Generative Adversarial Networks in Fog Environment. *IEEE Internet of Things J.*, 8(8):6247–6256.

Li, D., Chen, D., Jin, B., Shi, L., Goh, J., and Ng, S.-K. (2019). MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *Springer Int. Conf. on Artif. Neural Netw.*, pages 703–716.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal Adversarial Perturbations. In *Proc. of the IEEE Conf. on Comput. Vision and Pattern Recognit. (CVPR)*.

Nisioti, A., Mylonas, A., Yoo, P. D., and Katos, V. (2018). From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods. *IEEE Commun. Surveys & Tut.*, 20(4):3369–3388.

O'Shea, T. J. and West, N. (2016). Radio Machine Learning Dataset Generation with GNU Radio. *Proc. of the 6th GNU Radio Conf.*

O'Shea, T. J., Corgan, J., and Clancy, T. C. (2016). Convolutional radio modulation recognition networks. In *Int. Conf. on Eng. Appl. of Neural Networks*, pages 213–226. Springer.

Pourranjbar, A., Elleuch, I., Landry-pellerin, S., and Kaddoum, G. (2023). Defense and Offence Strategies for Tactical Wireless Networks Using Recurrent Neural Networks. *IEEE Trans. on Veh. Technol.*, pages 1–6.

Pourranjbar, A., Kaddoum, G., and Saad, W. (2022). Recurrent Neural Network-based Anti-jamming Framework for Defense Against Multiple Jamming Policies. *IEEE Internet of Things J.*, pages 1–1.

Sadeghi, M. and Larsson, E. G. (2019). Adversarial Attacks on Deep-Learning Based Radio Signal Classification. *IEEE Wireless Commun. Lett.*, 8(1):213–216.

Yuan, X., He, P., Zhu, Q., and Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. on Neural Netw. and Learn. Syst.*, 30(9):2805–2824.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., and Chandrasekhar, V. (2018). Adversarially Learned Anomaly Detection. In *IEEE Int. Conf. on Data Mining (ICDM)*, pages 727–736.

Zhang, L., Lambotharan, S., Zheng, G., AsSadhan, B., and Roli, F. (2021). Countermeasures Against Adversarial Examples in Radio Signal Classification. *IEEE Wireless Commun. Lett.*, 10(8):1830–1834.

Zhang, L., Lambotharan, S., Zheng, G., Liao, G., Demontis, A., and Roli, F. (2022). A Hybrid Training-Time and Run-Time Defense Against Adversarial Attacks in Modulation Classification. *IEEE Wireless Commun. Lett.*, 11(6):1161–1165.