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 permitem 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 nas áreas 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, 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

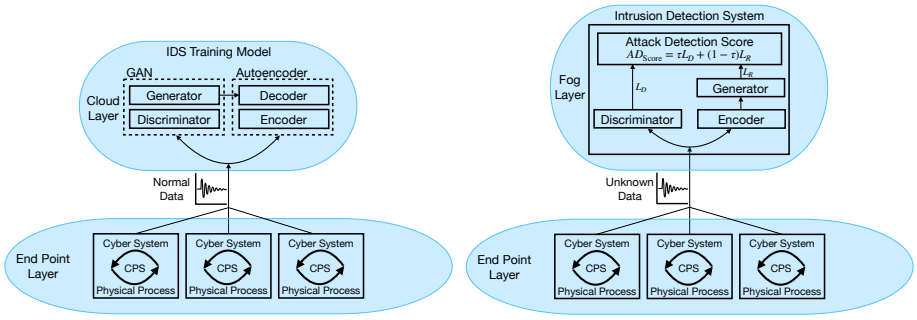
In this thesis, we advance the state-of-the-art in the security field by considering the cyber-attacks and adversarial attacks problems. Our contributions are divided in two parts. The first part concerns the use of ML for intrusion detection with the proposal of unsupervised IDSs to accomplish our first two specific objectives. It comprehends the contributions in [J4] and [J2] 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. Please find the complete list of our thesis' accomplishments, such as publications, patents, and awards in Section 4.

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. Figures 1a and 1b exhibit the architecture of the proposed IDS training and detection models, respectively. 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.



(a) Training Model **(b)** Detection Model

Figure 1. Proposed FID-GAN System Model (from [J4])

Our experiments show that our proposed IDS achieves detection rates that are higher than those of other state-of-the-art IDSs, while being faster than them due to our proposed Encoder. 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.

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. 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.

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. 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 accuracies or shorter detection times.

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. 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 2 shows the training model of our proposed adversarial attack technique.

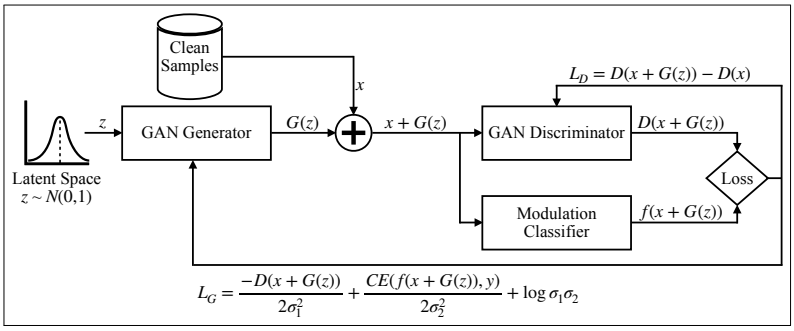


Figure 2. Our proposed training model (from [J3])

Our experiments show that it is possible to quickly craft small imperceptible perturbations that severely compromise modulation classifiers' accuracy and hence wireless receivers' performance. Therefore, it is urgently necessary to enhance deep learning-based modulation classifiers' robustness against adversarial attacks.

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 classifier that has been trained using noisy and adversarial samples to make it resistant to sample variation. 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.

Our experiments show that our proposed defense significantly diminishes the accuracy reduction caused by adversarial attacks on modulation classifiers, and outperforms other protection techniques by at least 18 percentage points.

4. Research Accomplishments

4.1. Grants and Awards

Our Ph.D. research has been recognized with the following grants and awards:

- **Microsoft Research Ph.D. Fellowship** (2022)
  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.
- Fonds de recherche du Québec - B2X Scholarship (2021-2022)
  First place in the 2021-2022 FRQNT's B2X Doctoral Scholarship competition.
- Nomination to the Vanier Canada Graduate Scholarship (2020-2021)
- Mitacs Accelerate Fellowship (2020-2021)

4.2. Publications

Our Ph.D. research contributed to the following published research articles.

4.2.1. Main contributions

J1 **P. F. de Araujo-Filho**, G. Kaddoum, M. Chiheb Ben Nasr, H. F. Arcoverde and D. R. Campelo, "Defending Wireless Receivers Against Adversarial Attacks on Modulation Classifiers," in IEEE Internet of Things J., vol. 10, no. 21, pp. 19153-19162, Nov. 1, 2023.

J2 **P. F. de Araujo-Filho**, M. Naili, G. Kaddoum, E. T. Fapi and Z. Zhu, "Unsupervised GAN-Based Intrusion Detection System Using Temporal Convolutional Networks and Self-Attention," in IEEE Trans. on Netw. and Service Manage., pp. 1–1, 2023.

J3 **P. Freitas de Araujo-Filho**, G. Kaddoum, M. Naili, E. T. Fapi and Z. Zhu, "Multi-Objective GAN-Based Adversarial Attack Technique for Modulation Classifiers," in IEEE Commun. Lett., vol. 26, no. 7, pp. 1583-1587, July 2022.

J4 **P. Freitas de Araujo-Filho**, G. Kaddoum, D. R. Campelo, A. Gondim Santos, D. Macêdo and C. Zanchettin, "Intrusion Detection for Cyber–Physical Systems Using Generative Adversarial Networks in Fog Environment," in IEEE Internet of Things J., vol. 8, no. 8, pp. 6247-6256, April 15, 2021.

J5 **P. Freitas De Araujo-Filho**, A. J. Pinheiro, G. Kaddoum, D. R. Campelo and F. L. Soares, "An Efficient Intrusion Prevention System for CAN: Hindering Cyber-Attacks With a Low-Cost Platform," in IEEE Access, vol. 9, pp. 166855-166869, 2021.

4.2.2. Collaborations

J6 P. Illy, G. Kaddoum, **P. F. de Araujo-Filho**, K. Kaur and S. Garg, "A Hybrid Multistage DNN-Based Collaborative IDPS for High-Risk Smart Factory Networks," in IEEE Trans. on Netw. and Service Manage., vol. 19, no. 4, pp. 4273-4283, Dec. 2022.

J7 M. -T. Nguyen, G. Kaddoum, B. Selim, K. V. Srinivas and **P. F. de Araujo-Filho**, "Deep Unfolding Network for PAPR Reduction in Multicarrier OFDM Systems," in IEEE Commun. Lett., vol. 26, no. 11, pp. 2616-2620, Nov. 2022.

J8 A. J. Pinheiro, **P. Freitas de Araujo-Filho**, J. de M. Bezerra and D. R. Campelo, "Adaptive Packet Padding Approach for Smart Home Networks: A Tradeoff Between Privacy and Performance," in IEEE Internet of Things J., vol. 8, no. 5, pp. 3930-3938, March 1, 2021.

C1 L. Luz, **P. Araujo-Filho**, and D. Campelo. "Multi-Criteria Optimized Deep Learning-based Intrusion Detection System for Detecting Cyberattacks in Automotive Ethernet Networks", in Anais do XLI SBRC, Brasília/DF, 2023, pp. 197-210.

C2 P. do Carmo, **P. Freitas de Araujo-Filho**, D. R. Campelo, E. Freitas, D. Sadok, "Machine Learning-Based Intrusion Detection System for Automotive Ethernet: Detecting Cyber-Attacks with a Low-Cost Platform", in Anais do XL SBRC, Fortaleza, 2022.

C3 L. Prado D'Andrada, **P. Freitas de Araujo-Filh**o, and D. R. Campelo, "A Real-time Anomaly-based Intrusion Detection System for Automotive Controller Area Networks", in Anais do XXXVIII SBRC, Rio de Janeiro, 2020, pp. 658-671.

4.3. Patents

Our Ph.D. research contributed to the following patents.

P1 M. Naili, **P. F. de Araujo-Filho**, G. Kaddoum and E. T. Fapi, "Detecting Anomalous Behaviour in an Edge Communication Network," World Intellectual Property Organization International Publication Number WO 2023/285864 A1, 2023.

P2 **P. F. de Araujo-Filho**, G. Kaddoum, M. Naili, E. T. Fapi and Z. Zhu, "Unsupervised GAN-Based Intrusion Detection System using Temporal Convolutional Networks, Self-Attention, and Transformers," World Intellectual Property Organization International Publication Number WO 2022/259125 A1, 2022.

4.4. Software Registration

Our Ph.D. research contributed to the following software registry.

(2020) Certificado de Registro de Programa de Computador. Processo No: BR512021001518-5, "Intrusion Detection for Cyber-Physical Systems using Generative Adversarial Networks in Fog Environment"

4.5. Publications for the general public

Our Ph.D. research has also been disseminated to the general public through the publication of several blog posts in https://substance.etsmtl.ca and https://www.sidechannel.blog.

5. 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

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.

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.

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.