Towards Evaluating Adversarial Attacks Robustness in Wireless Communication

Asmaa FTAIMI[1], Tomader MAZRI[2]
Laboratory of Advanced Systems Engineering
Ibn Tofail Science University
Kenitra, Morocco

***Abstract*—The emerging new technologies, such as autonomous vehicles, augmented reality, IoT, and other aspects that are revolutionising our world today, have highlighted new requirements that wireless communications must fulfil. Wireless communications are expected to have a high optimisation capability, efficient detection ability, and prediction flexibility to meet today's cutting-edge telecommunications technologies' challenges and constraints. In this regard, the integration of deep learning models in wireless communications appears to be extremely promising. However, the study of deep learning models has exhibited inherent vulnerabilities that attackers could harness to compromise wireless communication systems. The examination of these vulnerabilities and the evaluation of the attacks leveraging them remains essential. Therefore, this paper's main objective is to address the alignment of security studies of deep learning models with wireless communications' specific requirements, thereby proposing a pattern for assessing adversarial attacks targeting deep learning models embedded in wireless communications.**

***Keywords*—Adversarial attacks; deep learning; wireless communication; security; robustness; vulnerability; threat**

I. Introduction

Wireless communication represents an exciting and evolving field of study. It has significantly contributed to the development of telecommunications and had been at the source of LTE and 5G network implementations, and it continues today to lead advancement in 6G generation development. Nevertheless, this field of study has recently encountered several challenges to meet emerging telecom technologies' requirements. The necessity of optimisation and adaptability is crucial today to guarantee highly efficient wireless communications [1]. In this context, the researchers have immediately turned to the rapidly growing techniques of artificial intelligence, especially the deep learning models that have received considerable interest for their reliability in computer vision and object detection.

Deep Learning models have revolutionised many fields of study. Their expressivity and generalisation potentials have shown impressive outcomes in several areas, including wireless communications [2]. The latter have utilised deep learning models in the radio frequency spectrum and have harnessed their adaptability and flexibility to improve the wireless communication capacity. An obvious application of Deep Learning in Wireless Communications is spectrum estimation and detection and modulation classification. These two functionalities contribute significantly to enhancing the transmission quality while handling the channel effects encountered at the receiver [3,4].

However, several research studies have revealed that Deep Learning models contain inherent vulnerabilities that an attacker could eventually harness to perform malicious actions compromising wireless communications systems. Nevertheless, researchers usually opt for different hypotheses and follow different methods when testing adversarial attacks, rendering confronting, and comparing the latter challenging and complicated since the platform used in literature while testing adversarial attacks are not standardised. Whereas some researchers hypothesise a scenario where the attacker possesses complete knowledge of the system, known as white-box attacks, others deal with attacks where the attacker's knowledge is constrained, known as grey-box attacks. This disparity in assumptions provides a complex platform to make an accurate comparison of attacks on various algorithms.

Since no unified scheme has been developed for assessing attacks robustness, we have proposed a framework to analyse and evaluate the robustness of adversarial attacks in the context of wireless communication [27-30]. This paper aims to provide a standardised and unified platform for comparing different adversarial attack strategies against wireless communication systems. Through our study of adversarial attacks in literature, we have derived a list of criteria that we have considered to evaluate the complexity of the attack and its impact on the target system to obtain a global view of its robustness.

In this paper, we will extensively review the work devoted to studying the particularities of adversarial attacks targeting wireless communications to propose a pattern designed to evaluate the robustness of these attacks while incorporating the specifications related to this field of study.

First, we will introduce the applications of deep learning models in wireless communications. Then we will discuss the theoretical aspect of these attacks as well as the different factors involved in their identification and classification. Afterwards, we provide a review of the work highlighting certain particularities of wireless communications that could impact adversary attack success. Finally, we will elaborate on a pattern being proposed to evaluate the adversarial attacks' robustness in the context of wireless communications.

II. Related Works

A new research field has been conducted to extensively study the security of deep learning models and adversarial attacks leveraging their flaws [5]. Adversarial learning addresses the empirical evaluation of adversarial attacks by testing them in the physical world to experience them in a realistic scenario [6]. This research field is similarly focused on studying the theoretical aspect of these attacks by proposing a taxonomy for their classification and a threat model to describe the different aspects of the attack. However, the multiple suggestions carried out towards assessing adversarial attacks have mainly focused on attacks targeting computer vision or object detection models. Few works have addressed adversarial learning in the wireless communications context [27-30]. Indeed, the latter presents certain specificities to be considered in the study of attacks targeting the models they employ. Hence, it is necessary to adapt threat models and methods for evaluating adversarial attacks' robustness to wireless communication's technical and functional specifications.

III. Deep Learning Application in Wireless Communications

Wireless communications represent an essential field of study for developing networks and meeting innovative technologies' specific requirements. The evolution of telecommunications stems from this research field since it has brought LTE and 5G networks to the surface. It is also contributing significantly today to waveform design for 6G emerging networks. Wireless communications have been based on classical probabilistic and analytical methods. However, such an approach involves several limitations regarding channel modelling, interference handling, traffic management, error detection and correction, and security [7].

Wireless communications have evolved systems built on deep learning models to overcome the complexities encountered in earlier network generations. Wireless Communications leverage the expressiveness and capacity for generalisation of Deep Learning models toward addressing detection, classification, optimisation, and prediction problems and consequently guarantee quick, reliable, and secure communications.

A. Deep Learning for Communication Systems

The primary purpose of wireless communication is to ensure a message's reception in an optimal state by deploying resources efficiently. The transmission mechanisms deployed by wireless communications are handled through independent blocks, each dedicated to the specific functionality of the data transmission process, as shown in “Fig. 1”. Conventional approaches have focused on enhancing each block's functionality separately, thereby failing to achieve a proper optimisation of the overall system. Nevertheless, deep learning models flexibility currently provides the ability to address the optimisation needed for different blocks in parallel [8].

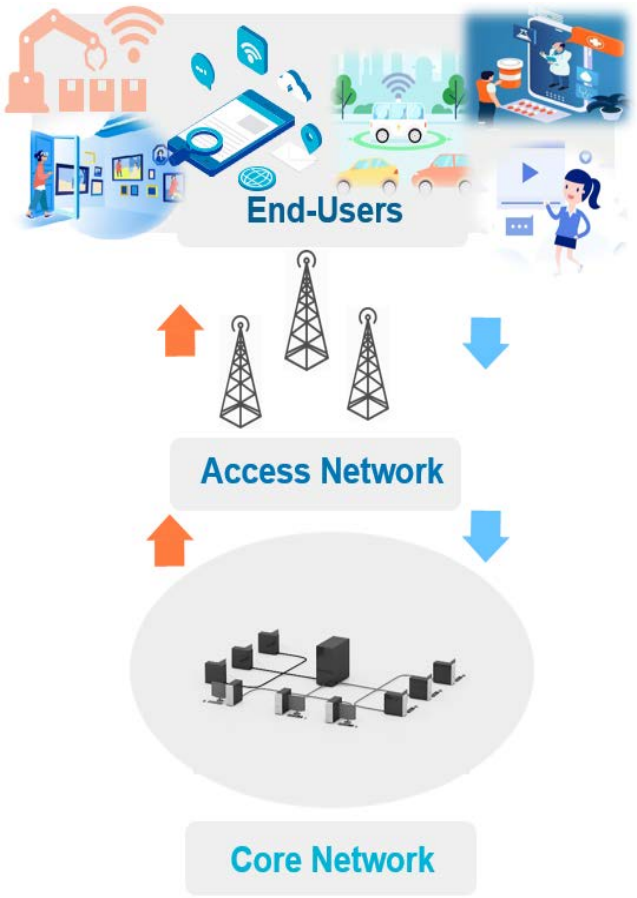

Fig. 1. Wireless Communication Architecture.

Moreover, Deep Learning in the end-to-end communication process has been employed significantly in implementing MIMO techniques. This technology involves integrating multiple antennas during the transmission and reception of signals to boost the spectrum's performance. Moreover, the implementation of multiple antennas appears to be computationally expensive and challenging in system optimisation. Thus, the use of Deep Learning in innovative studies [9,10] regarding the MIMO technique has overcome these challenges. The application of deep learning models has proven to be important when managing multiuser communication systems as well. These techniques could be applied to optimise the spectrum's exploitation for multiple users, yet they remained restrained by channel interferences.

Accordingly, an emerging technique known as Non-Orthogonal Multiple access NOMA has contributed to solving this trade-off by improving spectrum efficiency while minimising interference [11]. Currently, a new approach is also being considered for addressing these three issues in a synergistic scheme. It involves dealing with both emitter and receiver as one system designed as a single autoencoder [12], requiring a comprehensive optimisation.

B. Deep Learning for Spectrum Estimation, Detection, and Classification

Wireless communications have employed Deep Learning models' powerful capabilities in adaptability and flexibility to arm their systems with cognition. Cognitive radio could learn from collected data to adjust dynamically and rapidly the spectrum to performance and throughput demands. Among the crucial tasks that cognitive radios need to carry out, we cite signal detection and classification, although it is complex for classical feature-based algorithms that lack the flexibility to accommodate different types of signals. Deep learning model-based systems can overcome such problems since they have a high generalisation ability to classify and detect several types of signals, as shown in "Fig. 2".

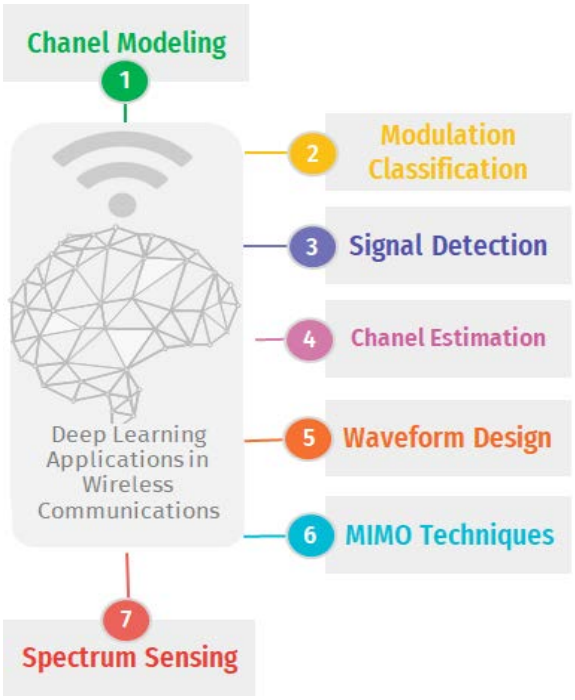

Fig. 2. Deep Learning Applications in Wireless Communications Systems.

1) Channel modeling and estimation: Channel modelling becomes essential to enhance the communication system's performance. For instance, autoencoders require a training phase approaching as closely as possible the real channel conditions. However, because of the channel effect, performing an efficient autoencoder training enfold several complexities. Hence reliance on GAN [13] to approximate interferences, noise, and multi-path effects to depict a channel model representing accurate and realistic behaviour [14].

2) Signal detection: Signal detection and classification is crucial functionality in wireless communications. It allows the control of system components and provides an up-to-date overview of the communications and events occurring in the system. Indeed, it ensures reliable detection of spectrum users and arising events, such as identifying interference sources for immediate response. Nevertheless, the spectrum is often shared for multiple simultaneous applications (TV, GSM, LTE, Radar, Etc.), which is challenging when identifying the wide variety of waveforms used. The conventional general and specialised detection methods lack scalability and depend on the SNR for signal detection and classification. Therefore, detection using these methods is difficult to perform, especially when the SNR signal to noise ratio is low [15].

3) Consequently, many studies consider employing CNN models that have proven their high performance in object detection and recognition, specifically in computer vision. O'Shea et al. [16] have examined the application of CNN models for signal detection and classification in the radio frequency spectrum. Their study utilised Gradient-Weighted Class Activation Mapping (Grad-Cam) for spectral event localisation and have achieved high performing results.

4) Modulation classification: O'Shea et al. [17] have evaluated the performance of CNN models in modulation classification by experimenting with channel effects such as multi-path fading to test the accuracy rate obtained under real-world conditions. Following the study carried out on a dataset of 11 types of modulations often used in wireless communication, the results obtained by CNN models vastly exceeded those produced by SVM or Naive Bayes, even for all used SNR ratios.

Although deep learning delivers considerable potential advantages, most recent studies have shown that they contain many vulnerabilities that attackers can harness to perform malicious manipulations [18]. Many researchers have recently examined the security of deep learning models. Some have considered the practical aspect by testing these attacks in the physical world, particularly in computer vision and object detection, while others have focused on studying the theoretical aspect of adversarial attacks. In the following section, we will present the taxonomy of these attacks and the different classifications proposed for their study.

IV. Threat Model

Recently, several research studies have focused on scrutinising the security of Deep learning models. Experiments conducted in a variety of fields of study have shown that these models are vulnerable. Their high flexibility potential and their adaptability have contributed considerably to their weakness [19]. Today, several types of attacks have been tested to reveal the security flaws of deep learning models. Indeed, an attacker potentially poisons the model by introducing malicious data during the training phase to affect its behaviour to the new input data. The attacker could also severely compromise the model even in the prediction phase through carefully designed inputs to exploit its inherent flaws to mislead it into producing inaccurate results [20]. All these considerations have motivated researchers to explore deep learning models' security using conventional security approaches, especially the study of their confidentiality, integrity, and availability (CIA). In this regard, a group of researchers, namely Barreno et al. [21], have developed a taxonomy dedicated to the security of these models, providing a comprehensive classification of adversarial attacks by highlighting the opponent's goals, his knowledge of the targeted system, his capabilities and the strategy he may employ to carry out the attack as illustrated in "Fig. 3".

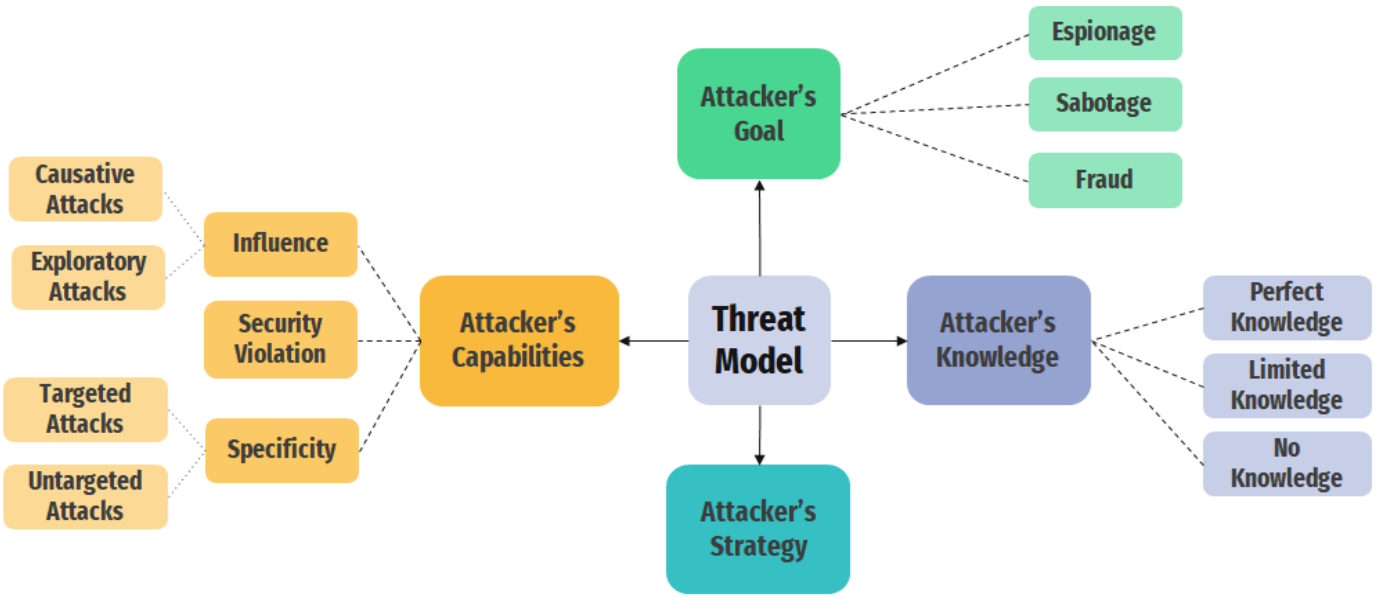

Fig. 3. Adversarial Learning Threat Model.

A. Attacker's Goal

While the same motivations might not necessarily guide attackers, their aims tend to converge around three main axes: espionage, sabotage, or fraud [22].

1) Espionage: In this context, the attacker seeks to derive sensitive information by exploiting vulnerabilities in the system. The leakage of sensitive information can compromise the confidentiality and the privacy of the system since the received information can be utilised to plan more advanced attacks and consequently cause very severe incidents.

2) Sabotage: The attacker can obstruct the system by either disabling important functionalities or by denying normal operations. Usually, this occurs when the adversary attempts to flood the model with incorrectly classified examples to increase the working time on false positive, or he can just as easily overload the system with a massive number of requests that require more computation time.

3) Fraud: Fraud in the Deep Learning system refers to the adversary's action of causing misclassifications or inaccurate predictions. In this case, the attacker takes advantage of existing vulnerabilities in the system to inject malicious inputs in the dataset or even modify the model's behaviour, and in this way, he causes severe damages.

B. Attacker's Knowledge

Regardless of the adversary's goals, the complexity of the attack he intends to carry out depends significantly on his knowledge of the targeted system. Papernot et al. [23] have elaborated a classification of the attacker's knowledge that can be represented in the following three levels:

1) Perfect knowledge level: in this scenario, the attacker knows everything about the model and the training dataset and can carry out white-box attacks. Nevertheless, this scenario remains unrealistic because it is almost impossible to have perfect knowledge about the target system.

2) Limited knowledge level: This scenario is more realistic and practical; nevertheless, it presents a range of possibilities: 1) Limited Knowledge attacks with surrogate model: when the attacker has limited knowledge of the model, he can use an alternative model with features similar to the targeted system's model's features in order to craft effective attacks, 2) Limited knowledge attacks with the surrogate dataset: when the adversary has no access to the training data he can use substitute dataset with similar characteristics to carry out efficiently the attacks.

3) Oracle or no knowledge: In this case, the attacker has no prior information about the training set, the model, or its features; he can perform black-box attacks.

C. Attacker's Capabilities

Besides factors seen in the previous sections, the opponent's means and potentials contribute largely to determining the attack's success [24]. The adversary's potentials could be categorised according to the following three main axes:

1) Influence: The attacker can compromise the targeted system using either a causative attack to introduce malicious data into the algorithm's training set or by exploiting the model's weaknesses by introducing specific inputs, often called adversary examples. The causative attack influences the model's behaviour contrary to the exploratory attack, where the adversary does not affect the model's behaviour [25].

2) Security violation: The security infraction committed by the adversary relies on the actions taken to compromise the targeted system. The attacker can cause integrity violations by crafting false-negative inputs that bypass the model without altering the usual tasks. However, the adversary causes an availability violation when he conducts false positives, leading

to a denial of service. Moreover, the adversary can also perform privacy violation attacks that aim to derive sensitive information about the users of the targeted system, the training dataset, or the features of the model [26].

3) Specificity: This characteristic determines how much the adversary is specific while performing the attack. Indeed, when the attacker has intentions to mislead the model for specific instances, he can perform targeted attacks, whereas, if he chooses to compromise the predictions or the classifications carried out by the model for a broad range of inputs, he must implement indiscriminate attacks [22].

D. Attacker's Strategy

An attack strategy can be elaborated by leveraging the model's vulnerabilities and flaws by considering the attacker's goals, knowledge of the targeted system, capabilities, and potentials. Therefore, the attack strategy in question is nothing more than an optimisation problem aiming to minimise the model performance by carefully crafting efficient and imperceptible perturbations to achieve its malicious objectives successfully [21].

V. Evaluating Adversarial Attacks in Wireless Communication

The study of adversarial learning in Wireless Communications presents new aspects beyond the scope of the examination of attacks targeting computer vision. Indeed, the adversarial attacks must be meticulously studied in this context. In the following paragraphs, we will examine the types of attacks targeting wireless communication systems. Then we will highlight important metrics that must be considered while assessing their robustness [27].

A. Type of Adversarial Attacks

Adversarial attacks in Wireless communications could be classified into two categories:

1) Direct Access Attacks (DAC): This category of attacks exploits the direct access to the classifier's input dataset to carry out malicious actions. The results obtained in [27] for this type of attack have shown that for symbol energy and jamming signal ratio Es/Ej of 30db, the FGSM attack produces a higher degradation than the one caused by the addition of AWGN Gaussian noise.

2) Over the Air Attacks (OTA): In computer vision, the attack reliability depends on selecting the perturbations that lead the model into misclassifications while remaining imperceptible to human eyes. Similarly to computer vision, Self-protect attacks are also interested in misleading the classifier while guaranteeing information transmission to the receiver with a defined modulation.

B. Metrics to Evaluate Adversarial Attacks in Wireless Communications

Besides, the evaluation of attacks and their success rates in Wireless Communications must adopt additional metrics aligned with signal transmission performance measures. Therefore, it is essential to consider the Bit Error Rate (BER) computation and the ratio of perturbing noise and modulated signal to estimate the opponent's attack's success rate, among other metrics.

1) Frequency offset: Before classifying the wideband signal, the systems initially identify the frequency of the signals and the time of transmission to convert these signals back to the baseband. Nevertheless, such operations may induce errors, as shown in [13], especially in centre frequency estimation, leading to frequency offset. The authors in [13] have noticed that raw-IQ-based AMC model accuracy dramatically decreases after adding frequency offsets. This encouraged Flowers et al. to examine frequency offsets' impact on adversarial attacks' success rate. The obtained results for 10 and 20 dB SNRs showed that even the most minor errors in frequency offset estimation could reduce the effect of adversary examples by increasing the model's accuracy by approximately 10%.

2) Time offset: To estimate transmission start and end times, the system employs an energy detection pattern that utilises a specific threshold of frequency power to determine the signal's existence in each instant. Incorrect evaluation of this threshold can result in false alarms or delays in the estimated transmission start time. In [28], the authors studied the impact of these parameters on the model accuracy rate. Indeed, in the absence of adversarial examples, the time offset does not significantly affect the model's accuracy rate. However, under adversarial conditions, they have noticed that translating the time-offset by four samples enhances the model accuracy rate by 20% for an Es/Ej ratio of 12 dB. Thus, they have concluded that the time offset can considerably reduce adversarial perturbations' impact on modulation classifier.

3) Multiple antenna usage: In [29], the authors have examined a wireless communication system in which the transmitter emits signals to receivers using several different modulation types. The receiver identifies the modulation types using a deep learning model classifier. In this context, an opponent can potentially introduce adversarial noise by employing multiple antennas to mislead the classifier and decrease its accuracy. They have demonstrated that using multiple antennas could enhance the opponent's attack robustness using a technique used in previous work [30] known as the maximum received perturbation power MRPP. They have evaluated this attack by emulating two different scenarios. In the first one, they have attacked while using adversaries operating in separate locations with only a single antenna. In the second scenario, they have performed the Elementwise Maximum Chanel Gain EMCG attack involving a single opponent yet with multiple antennas. These two scenarios have applied different techniques for power allocation. Kim et al. [29] also considered the opponent's attack on modulation classifiers while maintaining two essential requirements: 1) the perturbations introduced to the signal must be conceived to drive the targeted model to misclassify the modulations; 2) the power of the perturbations must not exceed maximum permissible levels so that they remain imperceptible to the receiver. The experimental findings indicated that the use of

multiple opponents would not degrade the classifier's performance significantly as a single adversary had employed the same antenna power [31]. The reason behind the obtained results is the lack of coordination and collaboration between multiple opponents in the second case since they do not focus on the same adversarial goal. The authors have demonstrated in their study the efficiency of EMCG attack. The EMCG attack associated with Gaussian noise gives the weakest results proving that Gaussian Noise's presence makes the adversarial perturbations detectable by the signal receiver.

VI. The Proposition of Assessment Pattern of Adversarial Attacks Robustness for Wireless Communication

Considering the results obtained in the study of adversarial attacks in the context of wireless communications, we will propose in this section a pattern to evaluate attack robustness, as shown in "Fig. 4". This pattern is inspired by our previous works [32] related to adversarial examples robustness assessment in computer vision. It is more adapted to the specificities of wireless communications.

The process is initiated by evaluating the attacker's knowledge of the targeted system. Attacks requiring complete knowledge of the targeted system are the least robust, while those designed with limited knowledge of the target system are proven to be the most robust. Afterwards, the attacker's potential and capabilities are analysed. Indeed, we have extensively detailed through Kim et al.'s [30] experience the impact of this factor in enhancing the robustness of the opponent's attack. In the previous section, we have already demonstrated that the adversary can improve the attack robustness by maximising the used antenna's power or increasing the number of antennas employed to carry out the adversarial attack.

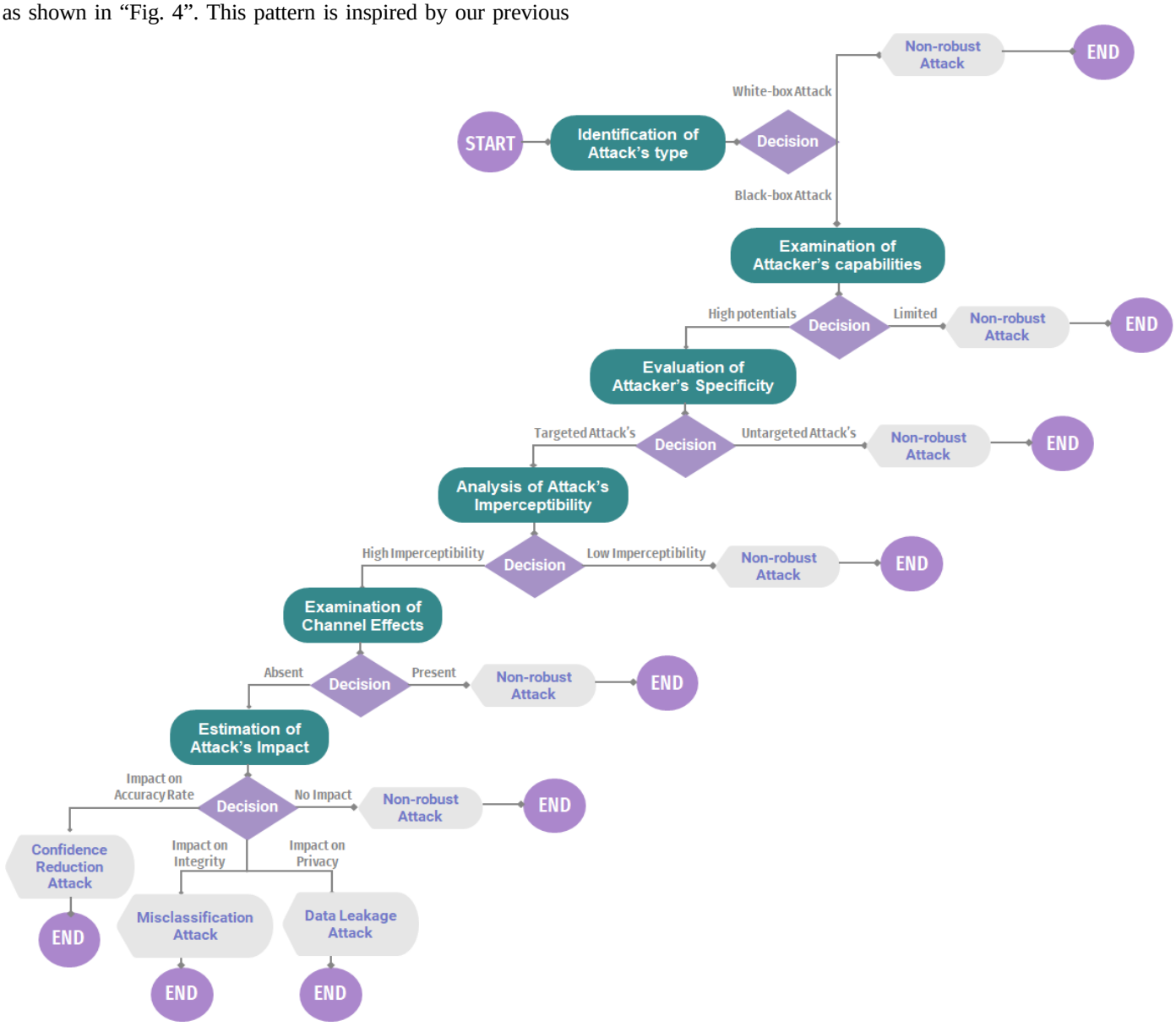

Fig. 4. Adversarial Attacks Robustness Evaluation Pattern.

Subsequently, the focus is on investigating the specificity of the attack. Attacks aiming at specific targets appear to be more robust than those seeking to introduce errors into the model without focusing on a specific target. Further, we analyse the imperceptibility of the perturbations introduced into the targeted system. Indeed, this parameter remains highly relevant to the attack's success rate. Further, we analyse the imperceptibility of the perturbations introduced into the targeted system. Indeed, this parameter remains highly relevant to the attack's success rate. It has been demonstrated in the studies conducted by [33] that as long as the crafted perturbations are perceptible, it is highly probable that they will be detected. This could be achieved in the context of wireless communications by keeping the power of the perturbations below the maximum permissible thresholds and thereby fooling the model without disrupting the signal transmission to the receiver.

In addition, the proposed pattern also involves examining the channel effects of Additive White Gaussian Noise, sample time offsets, and centre frequency offsets on the receiver. As in [34], researchers have shown that adding AWGN would significantly impact the adversarial examples compared to the model's accurate input data. In the presence of AWGN, the success rate of the adversarial attack is significantly reduced due to the identified sensitivity of adversarial examples to additive white Gaussian noise. Therefore, it is essential to include channel effects in the assessment process of adversarial attacks to match conditions that reflect a realistic scenario perfectly.

Finally, the pattern concludes with an estimation step of the attack's impact. To this end, we have proposed an analysis, including the adversary's objective and damaged components in the targeted system. Indeed, if the attacker fails to carry out malicious actions, we consider the attack with no impact on the target system. However, if it allows unauthorised access to the system by the adversary, then the attack has a significant impact on the confidentiality of the target system, yet its impact remains limited. Nevertheless, when the adversary carefully designs malicious perturbations to reduce model confidence in its predictions by diminishing its accuracy rate, then the impact is higher than the previous case. On the other hand, if crafted perturbations alter the model's output or influence its behaviour regarding the input dataset, then the impact is considered high since the attack affects target system integrity [35].

VII. Discussion

The approach we proposed in the previous section is designed to guarantee several advantages, including evaluating adversarial attack robustness according to several metrics, such as the type of attack, its specificity, its imperceptibility, and its impact. Indeed, we have considered analysing the adversary's knowledge and capabilities since they directly affect the attack's success rate. As far as the attacker is familiar with the components of the system, he can perform attacks causing tremendous damage. This feature allows the attacker to design imperceptible adversarial examples to carry out the attack and increase its specificity by targeting precise targets. In addition, We have included the attack's adaptability as an essential criterion for estimating its robustness. Indeed, the latter considerably affects the adopted attack strategy flexibility and therefore increases the challenge of the adopted defensive mechanism to mitigate the vulnerabilities of the targeted system. Moreover, we have conducted a comprehensive study of the attack's impact through an in-depth examination of the opponent's goal. Accordingly, we have developed a framework standardising the study and the assessment of different types of adversarial attacks. Our process provides the significant advantages of adaptability and generality since it can be applied to different models and can be tailored to different attacks and strategies that attackers may adopt in wireless communication systems.

VIII. Conclusion

Recently, growing attention in the scientific community has been dedicated to adversarial attacks. Throughout their studies, researchers have adopted several methods and established different hypotheses. This has made evaluating these attacks challenging since the platforms used in the experimentations are not standardised. Therefore, the proposed work suggests a unified method for evaluating adversarial attacks targeting deep learning models in wireless communications. This work has highlighted essential security aspects of the deep learning model used in wireless communications. In the different sections, we have explored the theory behind adversarial attacks as well as their practical application in the physical world. We have also examined the different studies carried out in this direction to draw a set of characteristics specific to wireless communications that greatly influence the success rate of the opponent's attack. At the end of this article, we have proposed a pattern devoted to evaluating adversarial attacks' robustness. Finally, the proposed model is designed to highlight the different characteristics of the attacks to provide an exhaustive evaluation that approximates the scenarios that can be encountered. Our future work will focus on implementing this framework using various Deep Learning models and different attacks to test its reliability in assessing adversarial attacks robustness in wireless communication systems.

References

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications," IEEE Journal on Selected Areas in Communications, 23(2), 201–220, 2005, doi:10.1109/JSAC.2004.839380.

[2] C. Clancy, J. Hecker, E. Stuntebeck, T. O'Shea, "Applications of Machine Learning to Cognitive Radio Networks," IEEE Wireless Communications, 14(4), 47–52, 2007, doi:10.1109/MWC.2007.4300983.

[3] S.M. Dudley, W.C. Headley, M. Lichtman, E.Y. Imana, X. Ma, M. Abdelbar, A. Padaki, A. Ullah, M.M. Sohul, T. Yang, J.H. Reed, "Practical Issues for Spectrum Management With Cognitive Radios," Proceedings of the IEEE, 102(3), 242–264, 2014, doi:10.1109/JPROC.2014.2298437.

[4] L.J. Wong, W.C. Headley, S. Andrews, R.M. Gerdes, A.J. Michaels, "Clustering Learned CNN Features from Raw I/Q Data for Emitter Identification," in MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM), IEEE, Los Angeles, CA: 26–33, 2018, doi:10.1109/MILCOM.2018.8599847.

[5] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, "Robustness May Be at Odds with Accuracy," ArXiv:1805.12152 [Cs, Stat], 2019.

[6] S. Qiu, Q. Liu, S. Zhou, C. Wu, “Review of Artificial Intelligence Adversarial Attack and Defense Technologies,” Applied Sciences, 9(5), 909, 2019, doi:10.3390/app9050909.

[7] T. Erpek, T.J. O’Shea, Y.E. Sagduyu, Y. Shi, T.C. Clancy, “Deep Learning for Wireless Communications,” ArXiv:2005.06068 [Cs], 2020.

[8] N. Rahimi, J. Maynor, B. Gupta, “Adversarial Machine Learning: Difficulties in Applying Machine Learning Existing Cybersecurity Systems,” 8.

[9] N. Samuel, T. Diskin, A. Wiesel, “Deep MIMO Detection,” ArXiv:1706.01151 [Cs, Math, Stat], 2017.

[10] H. He, C.-K. Wen, S. Jin, G.Y. Li, “A Model-Driven Deep Learning Network for MIMO Detection,” ArXiv:1809.09336 [Cs, Math], 2018.

[11] T. Erpek, S. Ulukus, Y.E. Sagduyu, “Interference Regime Enforcing Rate Maximization for Non-Orthogonal Multiple Access (NOMA),” in 2019 International Conference on Computing, Networking and Communications (ICNC), IEEE, Honolulu, HI, USA: 950–994, 2019, doi:10.1109/ICCNC.2019.8685624.

[12] I.J. Goodfellow, J. Shlens, C. Szegedy, “Explaining and Harnessing Adversarial Examples,” ArXiv:1412.6572 [Cs, Stat], 2015.

[13] T.J. O’Shea, T. Roy, N. West, B.C. Hilburn, “Physical Layer Communications System Design Over-the-Air Using Adversarial Networks,” ArXiv:1803.03145 [Cs, Eess], 2018.

[14] H. Ye, G.Y. Li, B.-H.F. Juang, K. Sivanesan, “Channel Agnostic End-to-End Learning based Communication Systems with Conditional GAN,” ArXiv:1807.00447 [Cs, Math], 2018.

[15] T. O’Shea, T. Roy, T.C. Clancy, “Learning robust general radio signal detection using computer vision methods,” in 2017 51st Asilomar Conference on Signals, Systems, and Computers, IEEE, Pacific Grove, CA, USA: 829–832, 2017, doi:10.1109/ACSSC.2017.8335463.

[16] T.J. O'Shea, T. Roy, T. Erpek, "Spectral detection and localisation of radio events with learned convolutional neural features," in 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, Kos, Greece: 331–335, 2017, doi:10.23919/EUSIPCO.2017.8081223.

[17] T.J. O’Shea, J. Corgan, T.C. Clancy, “Convolutional Radio Modulation Recognition Networks,” ArXiv:1602.04105 [Cs], 2016.

[18] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, “Intriguing properties of neural networks,” ArXiv:1312.6199 [Cs], 2014.

[19] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, B. Li, “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning,” ArXiv:1804.00308 [Cs], 2018.

[20] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, F. Roli, “Evasion Attacks against Machine Learning at Test Time,” ArXiv:1708.06131 [Cs], 7908, 387–402, 2013, doi:10.1007/978-3-642-40994-3_25.

[21] M. Barreno, B. Nelson, A.D. Joseph, J.D. Tygar, “The security of machine learning,” Machine Learning, 81(2), 121–148, 2010, doi:10.1007/s10994-010-5188-5.

[22] L. Muñoz-González, E.C. Lupu, The Security of Machine Learning Systems, Springer International Publishing, Cham: 47–79, 2019, doi:10.1007/978-3-319-98842-9_3.

[23] N. Papernot, P. McDaniel, A. Sinha, M. Wellman, “Towards the Science of Security and Privacy in Machine Learning,” ArXiv:1611.03814 [Cs], 2016.

[24] Z. Abaid, M.A. Kaafar, S. Jha, “Quantifying the impact of adversarial evasion attacks on machine learning based android malware classifiers,” in 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), IEEE, Cambridge, MA: 1–10, 2017, doi:10.1109/NCA.2017.8171381.

[25] L. Muñoz-González, E.C. Lupu, The Security of Machine Learning Systems, Springer International Publishing, Cham: 47–79, 2019, doi:10.1007/978-3-319-98842-9_3.

[26] L. Huang, A.D. Joseph, B. Nelson, B.I.P. Rubinstein, J.D. Tygar, “Adversarial Machine Learning,” 15.

[27] B. Flowers, R.M. Buehrer, W.C. Headley, “Evaluating Adversarial Evasion Attacks in the Context of Wireless Communications,” ArXiv:1903.01563 [Cs, Eess, Stat], 2019.

[28] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, “Universal Adversarial Perturbations,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI: 86–94, 2017, doi:10.1109/CVPR.2017.17.

[29] B. Kim, Y.E. Sagduyu, T. Erpek, K. Davaslioglu, S. Ulukus, “Adversarial Attacks with Multiple Antennas Against Deep Learning-Based Modulation Classifiers,” ArXiv:2007.16204 [Cs, Eess, Stat], 2020.

[30] B. Kim, Y.E. Sagduyu, K. Davaslioglu, T. Erpek, S. Ulukus, “Over-the-Air Adversarial Attacks on Deep Learning Based Modulation Classifier over Wireless Channels,” in 2020 54th Annual Conference on Information Sciences and Systems (CISS), IEEE, Princeton, NJ, USA: 1–6, 2020, doi:10.1109/CISS48834.2020.1570617416.

[31] B. Kim, Y.E. Sagduyu, K. Davaslioglu, T. Erpek, S. Ulukus, “Channel-Aware Adversarial Attacks Against Deep Learning-Based Wireless Signal Classifiers,” ArXiv:2005.05321 [Cs, Eess, Stat], 2020.

[32] A. Ftaimi, T. Mazri, “Evaluation and Analysis of Robustness of Adversarial Examples Attacks in Deep Neural Networks,” 6 in press.

[33] M. Sadeghi, E.G. Larsson, “Adversarial Attacks on Deep-Learning Based Radio Signal Classification,” ArXiv:1808.07713 [Cs, Eess, Math, Stat], 2018.

[34] J. Wang, J. Sun, P. Zhang, X. Wang, “Detecting Adversarial Samples for Deep Neural Networks through Mutation Testing,” ArXiv:1805.05010 [Cs, Stat], 2018.

[35] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, “Adversarial Attacks and Defences: A Survey,” ArXiv:1810.00069 [Cs, Stat], 2018.