Adversarial Machine Learning in Wireless Communications Using RF Data: A Review

Damilola Adesina, Chung-Chu Hsieh, Yalin E. Sagduyu, *Senior Member, IEEE*, and Lijun Qian, *Senior Member, IEEE*

***Abstract*—Machine learning (ML) provides effective means to learn from spectrum data and solve complex tasks involved in wireless communications. Supported by recent advances in computational resources and algorithmic designs, deep learning (DL) has found success in performing various wireless communication tasks such as signal recognition, spectrum sensing and waveform design. However, ML in general and DL in particular have been found vulnerable to manipulations thus giving rise to a field of study called adversarial machine learning (AML). Although AML has been extensively studied in other data domains such as computer vision and natural language processing, research for AML in the wireless communications domain is still in its early stage. This paper presents a comprehensive review of the latest research efforts focused on AML in wireless communications while accounting for the unique characteristics of wireless systems. First, the background of AML attacks on deep neural networks is discussed and a taxonomy of AML attack types is provided. Various methods of generating adversarial examples and attack mechanisms are also described. In addition, an holistic survey of existing research on AML attacks for various wireless communication problems as well as the corresponding defense mechanisms in the wireless domain are presented. Finally, as new attacks and defense techniques are developed, recent research trends and the overarching future outlook for AML in next-generation wireless communications are discussed.**

***Index Terms*—Adversarial machine learning, machine learning security, wireless security, wireless attacks, defenses.**

I. INTRODUCTION

MACHINE learning (ML) in general and deep learning (DL) in particular has found rich applications in various domains such as computer vision (CV) and natural language processing (NLP). Motivated by the success in those domains, there has been a growing recent interest in the development of artificial intelligence driven solutions for wireless communications using radio frequency (RF) data [1], [2], [3]. For example, convolutional neural network (CNN) models have been used for modulation recognition [4] and channel decoding [5]. In addition to CNN, feed-forward neural network (FNN) and Long short-term memory (LSTM) models have been used for device fingerprinting and identification [6], [7], [8], [9]. On the other hand, various deep neural network (DNN) models have been proposed for spectrum sensing and prediction [10], wireless resource management [11], beam prediction for initial access [12] and reconfigurable intelligent surfaces [13]. Although the application of DL-based techniques to wireless communications has shown promising results to solve complex problems for which analytical solutions are either unavailable or computationally infeasible, there are concerns in terms of reliability and robustness of such data-driven methods when compared to traditional signal processing methods in real-world deployments. One of such concerns is the presence of adversaries that may target the training process and/or testing (inference) process of DL.

With the increased adoption of DL, it is inevitable that adversaries will explore launching new attacks particularly on DL models, thus leading to a new paradigm called adversarial machine learning (AML). AML studies learning in the presence of adversaries with the goal of understanding the impact of AML attacks, defending against such attacks, and eventually developing secure DL systems. One type of AML attack (namely adversarial attack or evasion attack) occurs when an adversary crafts small perturbations to the original input of a neural network or develops thereby manipulating the operation of the DL system to cause an error in the inference process. These perturbations are not just white noise samples but they are specifically crafted to produce a vector in the input feature space that is capable of misleading the developed DL model. Classical examples of such attacks include adding small incremental value in the model's gradient direction with respect to the inputs or by solving a constrained optimization problem to produce a vector in the input feature space that is capable of misleading the target DL model.

AML attacks for CV and NLP have been extensively studied [14], [15], [16]. A canonical example is the misclassification of a panda image as gibbon because of the addition of a specifically crafted perturbation to the panda image so as to mislead the ML classifier [17]. Similarly, in the wireless domain, adversarial attacks can lead to misclassification of signals. As shown in Figure 1, the pre-trained ML model is a

Manuscript received 26 December 2020; revised 15 August 2021, 24 January 2022, and 13 June 2022; accepted 26 August 2022. Date of publication 12 September 2022; date of current version 24 February 2023. This work was supported in part by the U.S. National Science Foundation (NSF) under Award 2128482, and in part by the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Grant FA8750-15-2-0119 and Grant FA8750-15-2-0120. *(Corresponding author: Damilola Adesina.)*

Damilola Adesina and Lijun Qian are with the Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT Center), Department of Electrical and Computer Engineering, Prairie View A&M University, Prairie View, TX 77446 USA, and also with Texas A&M University System, College Station, TX 77840 USA (e-mail: dadesina@pvamu.edu; liqian@pvamu.edu).

Chung-Chu Hsieh is with the Center of Excellence in Cyber Security, Norfolk State University, Norfolk, VA 23504 USA (e-mail: ghsieh@nsu.edu).

Yalin E. Sagduyu is with the National Security Institute, Virginia Tech, Arlington, VA 22203 USA (e-mail: ysagduyu@vt.edu).

Digital Object Identifier 10.1109/COMST.2022.3205184

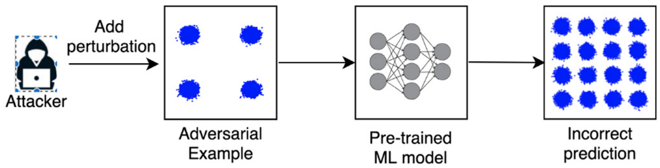

Fig. 1. An example of adversarial attack in wireless communications. An adversary adds perturbation to a test sample to create an adversarial example. Then, the pre-trained ML model wrongly predicts the test sample.

classifier for automatic modulation classification. An attacker launches an adversarial attack on the test data in the inference phase by adding a small perturbation to the original QPSK test data sample to form an adversarial example that causes the misclassification of a QPSK modulated signal as 16-QAM.

Data-driven, intelligent solutions empowered by ML/DL are considered as the key enablers in 5G and 6G wireless communications [18], [19], [20], [21]. As seen in [22], [23], the likelihood of jeopardizing ML/DL models deployed to enhance systems is a challenge that raise concerns for the safety of ML adoption. For instance, adversarial attacks causing incorrect ML/DL decisions may harm cyberphysical systems such as in autonomous vehicles [24], [25]. As such, to benefit from the gains of using ML/DL systems, there is an urgent need to ensure security and robustness of ML-based systems in the presence of adversarial attacks.

There have been multiple research efforts that aim to survey AML attacks as well as recent advances to detect and mitigate them in specific areas where ML/DL has found applications such as CV and NLP [26], [27]. Different from the existing reviews and surveys, this paper focuses on approaches to generate, detect and mitigate AML attacks in the wireless communications domain because of the unique characteristics of wireless communications and their effects on AML attacks. Such unique properties include the effect of the dynamic nature of the communication channel that introduces path loss [28], [29], [30], [31], [32], the effect of standardized wireless communication blocks to mitigate errors such as forward error correction (FEC) and hardware impairments [33] as well as access to data and heterogeneity in feature representation of the dataset [34], [35]. These unique properties discussed in Section III-D necessitate the review of the current attacks and mitigation methods of AML specifically for wireless communications as the methods presented in other domains such as CV and NLP may not be applicable. For instance, the adversary can directly query an API for CV applications without the need to send perturbations over a channel.

The design of DL algorithms for wireless communications must consider the new security issues such as identification and mitigation of adversarial signals [36]. Due to the dynamic properties of RF data, it is a non-trivial task to identify adversarial examples in RF data. DL methods rely on the premise that the training data represents the distribution of the underlying data generation process such that the trained model generalizes well on test data. However, this premise does not necessarily apply in the adversarial scenarios as these distributions change significantly due to adversarial effects that are not straightforward to anticipate or predict [37], [38], [39], [40]. With such adversarial attacks, wireless communications systems such as 5G and 6G with design components based on DL will be susceptible to disruptions, performance losses, and eventually failures. Therefore, it is critical to analyze the emerging wireless attacks due to AML, reduce and ultimately eliminate the impact of adversaries employing the AML attack techniques. An in-depth study of AML specifically for wireless communication systems is needed given the unique characteristics and emerging use cases of DL in such systems.

In order to address these critical needs, a comprehensive review of AML in wireless communications is provided in this paper. Because this topic is related to several important research topics including deep learning for wireless communications and adversarial machine learning, they are briefly covered in this review to provide some background and context for more in-depth discussions of AML in wireless communications. However, contrary to many existing surveys and reviews, this work primarily focuses on the review of adversarial attacks in wireless communications and adversarial examples in RF data and the corresponding detection and mitigation techniques. Interested readers may refer to many existing surveys and reviews on deep learning for wireless communications and adversarial machine learning for more detailed background information, as briefly discussed in Section I-A below.

A. Related Work

There are various surveys in related areas. For instance, [41], [42], [43], [44], [45], [46], [47], [48], [49], [50] surveyed attacks on wireless communications systems. The authors in [29], [33], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63] used various ML models for adversarial machine learning in wireless communications as seen in Table III while adversarial ML in other domains such as computer vision and NLP was discussed in [23], [25], [26], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73]. To the best of our knowledge, there is no thorough survey available for AML in wireless communication systems.

Traditional attacks on communication systems such as primary user emulation, jamming, eavesdropping attacks, and spectrum sensing data falsification have been extensively studied in [47], [74], [75], [76], [77]. To further secure communication systems, ML/DL techniques themselves can be used to detect and/or mitigate such traditional attacks [78], [79], [80]. On the other hand, AML has been reviewed and surveyed for various applications such as CV [26], autonomous vehicles [25], and cyber security [81]. The authors in [69], [82], [83], [84], [85] discussed the foundations of AML and provided detailed discussion on attack and defense strategies as well as relevant perspectives for designing more secure ML models. Reference [86] focused on the taxonomy of attacks to identify areas of action for researchers while [87] focused on adversarial example generation and defense mechanisms. Reference [73] provided a unified formulation for AML on graph data and compared attacks methods and defense strategies. Furthermore, [25], [64], [70], [71], [72], [88] provided various perspectives on AML in the CV domain. For instance [72] focused on AML in object recognition, [71]

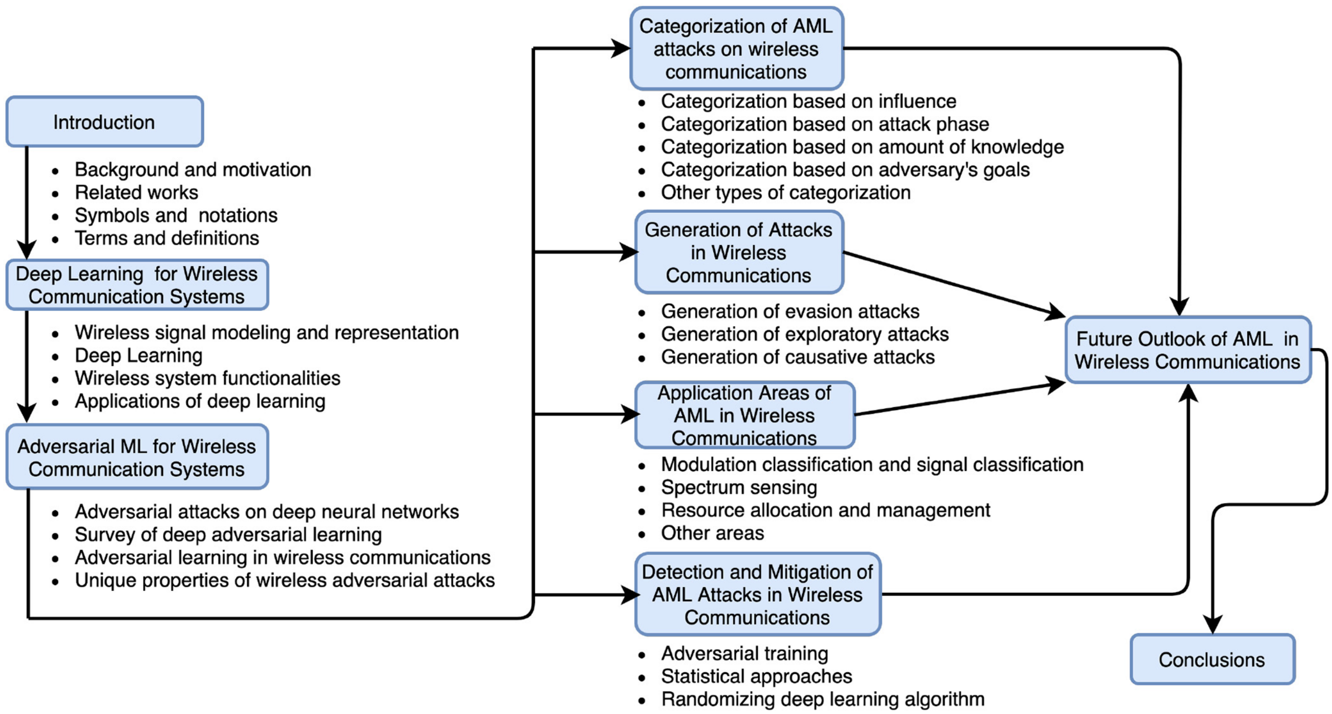

Fig. 2. Overview of the arrangement of the materials in this review.

focused on AML in image classification and [25] focused on AML in autonomous Vehicles.

These works presented a taxonomy of attacks, defense strategies and an outlook to the future. The scope of this review differs from previous works as it explores AML attacks that directly target the growing ML/DL applications in wireless communications. To the best of our knowledge, this is the first review work that focuses on AML in wireless communications by discussing the unique properties of wireless communications and its effects on AML. These AML attacks are effective because they operate with low spectrum footprint, and therefore, they are stealthy (hard to detect) and energy-efficient [24], [36], [56], [62], [89]. For instance, [52] showed that the DL-based exploratory attack is more effective in reducing the system throughput and more energy efficient than traditional random jamming attacks. The random jamming attack jams the transmission channel in some randomly selected instances while the DL-based exploratory attack jams in carefully selected time slots, thus using less energy. Due to the unique properties of wireless communications, it is essential to understand and quantify the impact of AML attacks on wireless communications and develop defense mechanisms that are tailored for the wireless domain.

B. Organization of Paper

An overview of the arrangement of the materials in this review is given in Figure 2 to help the readers navigate through this review. The symbols and notations as well as the definition of some important terms used in this review are summarized in Section II-A and Section I-C, respectively. Section II briefly reviews deep learning for wireless communication systems. Adversarial attacks and AML in wireless communications are reviewed in Section III. Section IV explores different types of AML attacks and provides a taxonomy of AML attacks in the wireless domain. Section V discusses the generation of AML attacks. Section VI reviews the application areas of AML for various wireless communication problems. Section VII presents the AML attack detection and mitigation methods. Section VIII discusses the future outlook of AML in wireless communications. Section IX concludes the paper.

C. Terms and Definitions

Terms and definitions used in this review are summarized in Table I-C.

Terms	Definitions
Adversary	An agent that adds perturbation to clean data with the aim of fooling ML models.
Adversarial Perturbation	The perturbation (noise or disturbance) signal generated by an adversary and added to clean data to alter it so as to fool an ML based system such as a classifier.
Adversarial Example	A sample of the data that has been altered by adding a perturbation to fool an ML based system.
Adversarial Attack	The process by which an adversary injects the adversarial example to undermine the ML-based system.
Adversarial Machine Learning	Learning in the presence of adversaries with the goal of understanding the impact of AML attacks, defending against such attacks, and eventually developing secure DL systems.
Adversarial Training	A defense mechanism where perturbed samples are added to the training data to boost model robustness against adversarial attacks.
Transferability	The ability of an adversarial example to retain its potency when transferred between different models.
Robustness	The ability of a model to remain reliable under changing conditions including adversarial attacks.

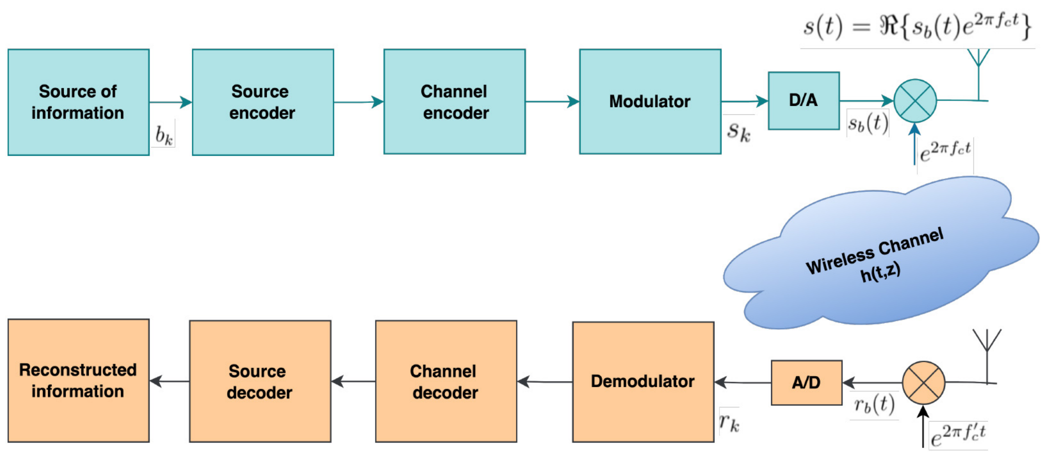

Fig. 3. Communication system diagram.

II. Deep Learning for Wireless Communication Systems

A. Symbols and Notations

X	total data points
x	a data point
x^*	adversarial example
Y	class (label) for dataset
y	class for a data point
y'	particular target label
r_x	adversarial perturbation
w	model parameters
J	cost function
α	step size
ϵ	small positive number
C	number of classes
$\bigtriangledown$	gradient operator
f_w	neural network classifier function

There is a lot of interest in developing artificial intelligence for wireless communication systems using radio frequency (RF) data. This is because the ability to learn and intelligently respond to dynamic and complex operating conditions will be of utmost importance in future wireless systems. It is envisioned that the knowledge of current operating conditions and environment leveraging on wireless big data analytics will allow communication systems to make the best opportunistic decisions [19], [90], [91], [92]. To achieve this, work is done to develop efficient algorithms and methods to extract meaningful information from complex and massive datasets from communication systems. Research in DL for wireless communications have gained tremendous attention in recent times, particularly the application of machine learning and deep learning to solve wireless communication problems using RF data. In this section, we present the foundation of wireless signal model, the deep neural network and the application of DL in wireless communication systems.

B. Wireless Signal Modeling and Representation

Traditional wireless communication systems are primarily comprised of a transmitter, a channel and a receiver. A block-based design principle is pursued to split the communication systems into multiple independent blocks. The functional blocks that are typically optimized with the aid of mathematical model and expert knowledge include source encoder (which compresses the input data and removes redundancy), channel encoder (which ensures the system can cope with the effects of the channel by adding redundancy to the source encoder output) and the modulator (which provides a mapping based on the desired received signal level and data rate) [1], [93], [94], [95]. The sequence of source information bits b_k is transformed by coding to a binary sequence mapped to symbols s_k through modulation as shown in Figure 3. The modulated signal is passed through a pulse filter. The resulting signal is reconstructed to a continuous analog time signal $s_b(t)$ by passing it through a digital-to-analog (D/A) converter. $s_b(t)$ is the baseband signal shifted by the carrier frequency f_c to obtain the bandpass signal $s(t) = \Re\{s_b(t)e^{2\pi f_c t}\}$ [95] expressed as:

$$s(t) = \Re\{s_b(t)\}\cos(2\pi f_c t) - \Im\{s_b(t)\}\sin(2\pi f_c t). \quad (1)$$

At the channel, the transmitted signal encounters impairments such as timing drift, frequency offset, noise. The variations in the channel can be modeled as large-scale fading and small-scale fading. The received signal $r(t)$ is the corrupted form of the signal transmitted due to hardware imperfections and channel impairments. The effect of the channel is modeled as a finite impulse response (FIR) filter $h(t,\tau)$. Therefore, the received signal $r(t)$ is the convolution of the signal $s(t)$ and the channel $h(t,\tau)$ given by $s(t) * h(t,\tau)$. $r(t) = \Re\{r_b(t)e^{2\pi f_c' t}\}$, where $r_b t$ is the baseband complex envelope described by:

$$r_b(t) = (s_b(t)h_b(t,\tau))\frac{1}{2}e^{j2\pi(f_c - f_c')t + \varphi(t)} + n(t). \quad (2)$$

The received signal $r(t)$ is passed through an amplifier and low-pass filter. The resulting continuous-time signal is converted to its digital version by an analog to digital converter. This is achieved by sampling it at rate $f_s = \frac{1}{T_s}$ samples per second, giving us a discrete representation of the signal made up of the in-phase and the quadrature components and representing the complex raw time-series signals r_k. The

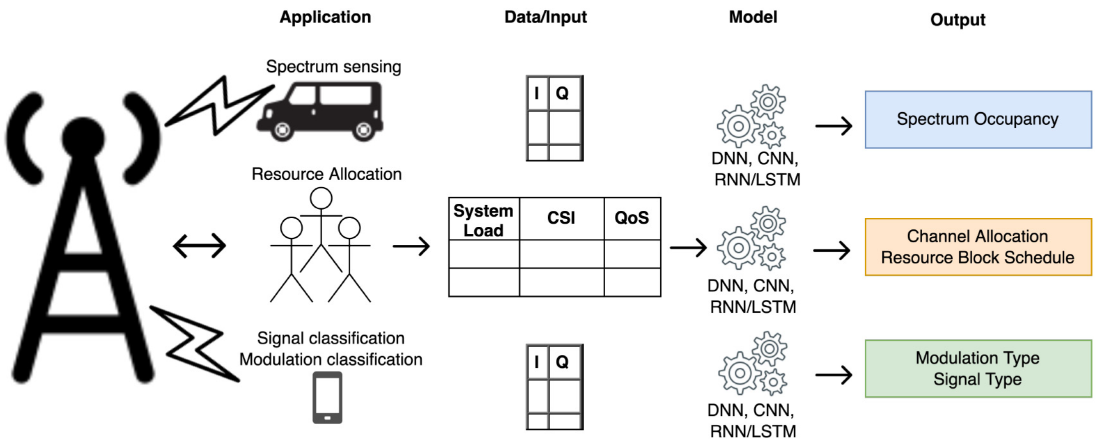

Fig. 4. System model for the applications of DL in wireless communications.

signal is demodulated and passed through the decoder to obtain the reconstructed information [95].

C. Deep Learning

Deep learning can automatically extract relevant features from data with no need for hand-engineering, thus making it more beneficial in most cases than traditional machine learning algorithms. The main types of deep learning models are:

1) Deep Neural Network (DNN): a type of artificial neural network (ANN) with multiple hidden layers that can extract and transform features automatically [96]. It is also considered a universal function approximator that can identify inherent patterns and discover linear and non-linear associations between features. It is computationally and memory intensive for training due to its large number of parameters. However, recent advances in the field, such as the development of tensor processing units (TPUs), have reduced computation time. A single layer of a neural network is formulated as $y = g(w^T x + b)$ where y is the output of the layer, w are the weights, x is the input data, b is a bias term and g is the activation function.

2) Convolutional Neural Network (CNN): a type of feed-forward neural network known to perform convolution operations described in Equation (3) for a multi-dimensional data X as input and a multi-dimensional filter K. The layers identify local patterns by using the same geometric transformations (filters) to various spatial locations in the input data. The convolution operation is achieved by taking the dot product of the filter K with dimension m by n and the input data across the input space to get a feature map S of dimension i by j as described in Equation (3). CNNs are well suited for a dataset with a grid-like topology such as images and time-series data. They leverage parameter sharing and sparse interactions to improve efficiency and reduce the memory requirement.

$$S(i,j) = (K * X)(i,j) = \sum_m \sum_n X(i-m, j-n)K(m,n) \tag{3}$$

3) Recurrent Neural Network (RNN): a type of artificial neural network optimized for sequential input tasks, such as speech recognition, as past observations influence its output. This is because it can learn from long-term dependencies and remember past observations. The dependence of the current output on past output indicates that RNN shares the same weights across several time steps. The dependency is in the form of a feedback loop connected to the output from the past, which distinguishes it from other types of neural networks [96].

DL is primarily using a neural network with multiple layers that perform non-linear transformations on data to provide some output. For example, supervised learning using a DNN is expressed as a mapping f that models the mapping from the input data to a class (label). To learn this input-output relationship, a weight matrix w is estimated. A loss function $l(x, y, w)$ is computed as a point-wise measure of error between the model prediction $\hat{y} = f(x)$ and the observed ground truth y, where $x \in X$ is a sample data point in data X and $y \in \mathcal{C}$ is a class in $\mathcal{C}$. To estimate w, a cost function $J(w)$, which is the average loss over all (n) training data samples in X, is computed as

$$J(w) \equiv J(X, y, w) = \frac{1}{n} \sum_{(x_i, y_i)} l(x_i, y_i, w). \tag{4}$$

The model is derived by minimizing the cost function $J(w)$ as

$$\operatorname*{argmin}_{w \epsilon \mathbb{R}^n} J(w). \tag{5}$$

Then, the DNN classifier indexed by the DNN weights w is determined as $f_w : X \rightarrow \mathbb{R}^C$, where C is the number of classes in $\mathcal{C}$. Interested readers may obtain details of DL in various sources such as the book [96].

D. Wireless System Functionalities and Applications of Deep Learning

DL has been extensively used in wireless communications as extensively surveyed before [1], [2], [3], [90], [93], [97], [98], [99], [100], [101], [102], [103]. These previous studies highlight various models and applications of DL in wireless communication systems some of which are shown in Figure 4. The authors in [1] gave an overview of deep learning for wireless communications focusing on physical layer

TABLE I
DEEP LEARNING APPLIED TO WIRELESS SYSTEM FUNCTIONALITIES

System functions	Input	Output	DL Models Applied
Spectrum sensing	I/Q signal	Spectrum occupancy	DNN, CNN, RNN/LSTM
Signal classification	I/Q signal	Type of signals or users	DNN, CNN, RNN/LSTM
Modulation classification	I/Q signal	Type of modulations(e.g., QPSK, QAM, etc.)	DNN, CNN, RNN/LSTM
Resource allocation	Channel state information, system load, users QoS requirements, etc.	Channel allocation, resource block schedule, etc.	DNN, CNN, RNN/LSTM

communication, spectrum situation awareness and wireless security in deep learning applications. We discuss some of the existing works in various application areas including modulation classification, spectrum sensing, resource allocation and protocol design.

- *Modulation and signal classification:* DL has been applied to identify the modulation type or signal in received signals. The modulation classification problem is an extensively studied problem in the wireless communication domain using DL. While modulation classification seeks to recognize the modulation scheme of a detected signal to aid correct demodulation of the received signal, a signal classification algorithm can help in spectrum monitoring, spectrum management, and secure communications. Numerous approaches have been used to solve this problem using DL. These approaches include leveraging the constellation density of the signal for classification [104], [105], [106] and building DL models that use RF data for classification [107], [108], [109]. While [1] focused on DL for physical layer communications, [110] provided a survey of DL-based modulation classification with a focus on data representation and signal preprocessing and [111] surveyed modulation recognition with a focus on the various types of DL algorithms used.
- *Spectrum sensing:* The objective of spectrum sensing is to detect the availability of frequency channels in dynamic spectrum access enabling technology such as cognitive radios. In recent times, machine learning-based sensing has been proposed in previous works [112], [113], [114], [115], [116], [117], [118]. The spectrum sensing application is typically developed as a classification problem that predicts if the channel is occupied or free. Figure 4 shows spectrum sensing in a vehicular network. Spectrum sensing applications have been surveyed in [119], [120].
- *Resource allocation and management:* In wireless communications, resource allocation and management such as beamformer design and power control plays a central role in the efficient operation of the communication system. With the significant increase in the interaction of different end users, communication technologies, and network operators as indicated in Figure 4, the use efficient resource allocation and management techniques such as DL-based approaches becomes inevitable. Before now, numerical optimization algorithms which are often complex are used for resource allocation and management. Numerous DL-based algorithms are being developed to make resource allocation more efficient [11], [121], [122], [123], [124], [125]. A large dataset is collected to train and develop DL models for various resource allocation and management problems. The work by [123] assigned resources to non-real time service in advance by learning the resource allocation pattern in a prediction window while DL is used to solve sub-band and power allocation problem in a multi-cell network [126]. The works in [122], [127] first learn the resource allocation schemes and prompt the appropriate action policy a priori to avoid and alleviate the congestion in an intelligent fashion. Resource allocation to enable coexistence of transceiver pairs and Wi-Fi users in LTE-U networks is proposed in [128].

More details on the current trends and application of machine learning in wireless communications systems are presented in [97], [99], [101], [129], [130], [131].

III. ADVERSARIAL MACHINE LEARNING FOR WIRELESS COMMUNICATION SYSTEMS

In this section, we discuss the deep adversarial learning starting with a brief introduction on adversarial attacks on deep neural networks and how adversarial attacks work. Then a survey of deep adversarial learning in other domains such as computer vision, graphs, and natural language processing is presented. Also, adversarial learning in wireless communications and how the unique properties of wireless communication systems affects adversarial attacks are discussed.

A. Adversarial Attack on Deep Neural Network

As an example of AML attacks, we describe next how the evasion or the so-called adversarial attack works. We will discuss other types of AML attacks in Section IV. For every $x \in X$, $f_w(x)$ assigns a label of the form $\hat{y} = \text{argmax}_{y \in \mathcal{C}} f_w(x, y)$, where $f_w(x, y)$ corresponds to the output of f_w for class y in $\mathcal{C}$. An adversarial attack on the classifier f_w aims to modify the classified input x with perturbation r_x such that $f_w(x + r_x) \neq f_w(x)$. For the stealth nature of this attack, it is required that the perturbation r_x should be small such that the difference of $x + r_x$ from the unperturbed input data x remains small. Typically, this requirement is relaxed to constraint $||r_x||_p \leq \epsilon$, where ϵ is a small positive number and $||\cdot||_p$ represents the l_p norm with $p \in \{1, 2, \infty\}$ denoting the type of norm. The adversarial perturbation r_x for x and f_w is determined by solving the following optimization problem:

$$\begin{aligned} &\min_{r_x} \quad ||r_x||_p \\ &s.t. \;\; f_w(x) \neq f_w(x + r_x). \end{aligned} \tag{6}$$

TABLE II
SURVEY OF DEEP ADVERSARIAL LEARNING IN DIFFERENT DOMAINS

Application area	Previous work
Computer Vision	[133]–[136]
Natural Language Processing	[137]–[144]
Graphs	[145]–[148]
Computational Biology	[149]–[152]
Electronic health records	[152]–[156]
Cancer Diagnosis	[157]–[162]

In the context of wireless communications, the l_2-norm is a natural choice for r_x as it accounts for the perturbation (signal) power [89], [132] compared to other norms such as the l_1-norm that has been used in the CV domain to deter the variation from human perception [89].

B. Survey of Deep Adversarial Learning

DL has found numerous applications in various areas and has provided major advancements to solving problems that have appeared unsolvable in the past. DL has achieved unprecedented results at scale in solving hard scientific problems. These advancements have been described in various surveys as detailed in Table II.

For instance in the computer vision domain, surveys have focused on applications of DL, including object detection [133], [134], [135], [136], 3D Data processing [163], [164], [165], [166], Hyperspectral Image Analysis [167], [168], [169], [170], [171], Dialogue Systems [137], [138], [139], [140] and speech processing [141], [142], [143], [144] in Natural Language Processing (NLP) and other areas such as Bioinformatics and Computational Biology [149], [150], [151], [152], Electronic Health Record (EHR) Analysis [152], [153], [154], [155], [156] and Cancer Diagnosis [157], [158], [159], [160], [161], [162].

Although DL has achieved remarkable success in various areas, DL is susceptible to attacks such as small perturbations which can significantly change the prediction of the DL algorithm as described in Section I. Next, we discuss the recent work in deep adversarial learning under computer vision, graphs and natural language processing.

- *Deep Adversarial Learning in Computer Vision:* Applications such as driver-less cars and surgical procedures have experienced significant growth with the advent of deep learning and its application. With these developments are the concerns of how such critical areas of life are impacted by the presence of adversaries especially due to the imperceptible nature of the attack from adversaries [26], [183], [184], [185], [186], [187], [188]. This became evident from [17] where a small imperceptible perturbation added to the input of image classification changes the label from stop sign to speed limit sign [189]. The realization of the potential vulnerabilities has led to the design and development of various algorithms to mitigate the effect of such perturbations and ensure model robustness. These include adversarial training to augment training data with adversarial examples in training phase [14], [17], [190], [191], [192], [193]. In building state-of-the-art models, more methods to mitigate the effect of adversaries in computer vision are surveyed in [26], [64], [82], [184], [194].
- *Deep Adversarial Learning in Graphs:* Graph-based learning is unique because it takes into consideration the input features and relationship of the features during training to account for connections between data samples. This is important because the data structures are not independent thus changing a connection or feature will change the prediction from others [195]. This form of learning has found applications in profiling social network users [196], [197] and analysis of biological interaction graphs [198]. The connection and dependence in the graph-based neural network could amplify the effect of perturbations when attacked. Adversarial learning is used in [145], [199], [200], [201] to investigate the robustness of graph models to various adversarial attacks using the ideas of graph convolution and also generate adversarial attacks that can easily be transferred to other models. Also, to mitigate the effect of attacks, adversarial training is used in [202] to design and implement a novel adversarial regularized framework that embeds the topological information and node content into a vector representation, from which a graph decoder is further built to reconstruct the input graph. A similar approach is used in [203] to embed each vertex in a graph in a low-dimensional vector space. Further studies on graph models and methods to mitigate attacks can be found in [73], [146], [204].
- *Deep Adversarial Learning in Natural Language Processing:* Applications such as speech recognition and text classification [205], [206] in the NLP domain also benefit from the growth in DL. An attacker can perturb the speech to fool the DL model of speech recognition systems such as Apple Siri. The work in [207] crafted audio adversarial examples on the state-of-the-art speech recognition system, Deep Speech [208], by adding inaudible sound perturbation. The attack works with 100% success. The work by [209] showed the effect of attacks on text classification. This can be in the form of letter, word, phrase manipulation or an attack on the word embedding [27], [210], [211]. Also, the adversarial sentences that have the same syntax structure as the original sentence and are grammatically correct were explored in [212], [213], [214]. This attack is achieved by using synonyms, adding words with different meaning in different contexts or paraphrasing. Adversarial training to mitigate the adversarial attack has also been proposed in NLP [215], [216], [217], [218]. Other methods such as randomized smoothing [216], supervised contrastive adversarial learning (SCAL) [218] and adversarial debiasing [219] are used in literature. Further surveys and reviews on adversarial learning in NLP include [215], [220], [221], [222].

C. Adversarial Learning in Wireless Communications

The application of DL to the wireless communications domain differs significantly from other application areas such

TABLE III
DL MODELS FOR ADVERSARIAL MACHINE LEARNING IN WIRELESS COMMUNICATIONS

Area	Sub-area	Papers
Signal related	signal classification	[28], [57], [59], [132], [172], [173] (CNN)
	modulation classification	[29] (DNN); [30]–[32], [35], [55], [174]–[176] (CNN)
	waveform	[56], [177] (CNN)
	rf fingerprinting	[32] (CNN); [178] (RNN)
Spectrum related	jamming	[51], [52] (DNN); [60], [61], [179](RL)
	spectrum deception	[58] (RL)
	spectrum poisoning	[51], [53] (DNN); [34] (CNN)
	perturbation	[54] (DNN, CNN)
	spectrum flooding attack	[63] (RL)
Wireless channel related	CSI estimation	[89] (CNN)
	forward error correction	[33] (CNN)
Privacy related	membership inference	[180], [181] (DNN)
	cooperative jammer	[182] (CNN)
	priority violation	[51] (DNN)
Resource allocation	radio access network slicing	[62] (RL)

TABLE IV
SUMMARY OF PREVIOUS WORK ON AML ATTACKS IN WIRELESS COMMUNICATIONS

Categorization	Attack Type	Description	Paper
Attack Type	Exploratory	Fathom the inner workings of model	[34], [51]–[53], [59], [60], [62], [63], [180], [228]–[230]
	Evasion	Fool ML models	[28]–[30], [33]–[35], [51], [53]–[55], [57], [58], [132], [174]–[176], [176], [178], [182], [231]
	Causative	Manipulate the training process	[34], [51], [53]
Adversary's Goal	Targeted	Cause specific errors	[28]–[30], [57], [59], [174], [182]
	Non-Targeted	Cause any errors	[28], [29], [32], [89], [174], [178]
Amount of Knowledge	White-box	Full knowledge	[28]–[30], [32], [54], [59], [89], [132], [174], [178], [182]
	Black-box	No knowledge	[28], [29], [32], [35], [54], [57], [61], [132], [179]–[181], [229]
	Gray-box	Limited knowledge	[61], [172]

as CV and NLP because specialized domain knowledge of wireless communications is needed for proper data representation, preprocessing, and result interpretation. In addition, channel, interference, and traffic effects shape both legitimate communications and others. These properties make the RF data unique and possibly have a different response to the adversarial attacks. Some unique properties of RF data are identified and discussed in relation to adversarial example generation and its effect in Section III-D.

Table III shows various areas in wireless communications where AML occur and various DL models used. These works establish the detrimental effect of adversarial attacks on wireless communication systems. The different types of attacks are categorized as detailed in Section IV and Table IV summarize the previous work on AML attacks in wireless communication detailing the attack type, categorization of attack, a description of the attack and the relevant publications.

The body of knowledge on adversarial learning in wireless communications is explored in subsequent sections of this paper that include the uniqueness of wireless adversarial attacks, categorization of AML attacks, attack generation methods, some application areas and AML attack detection and proposed mitigation schemes.

D. Unique Properties of Wireless Adversarial Attacks

DL has a strong potential to assist network decisions in achieving optimal resource management and future networks are envisioned to rely on DL-driven agents for functions such as network management automation and optimization of radio interfaces. The evaluation of the extent to which an adversary can affect such a complex system becomes critical in designing future communication systems with DL. This evaluation becomes critical as there are unique properties in the wireless communication domain as discussed in this section.

- *Effect of the communication channel:* The wireless communication channel has a major impact on adversarial attacks as they need to be launched over the air to reach the target receivers that host the victim DL models. One aspect is that the channel will impose path loss and phase on adversarial perturbations, and could weaken and/or change the direction of adversarial signals as they travel through wireless channels before reaching the target receivers. For an adversarial attack to succeed over the air, the adversary needs to account for the dynamic nature of channels when crafting the adversarial perturbations [28], [29], [30], [31], [32]. In addition, the adversary may not have access to the DL model at the target receiver and the process of data gathering by the adversary for training a surrogate model is typically performed through a channel, which implies that the training data used by the adversary is imperfect by default and the effectiveness of the surrogate model trained by the adversary strongly depends on channel effects [59]. Furthermore, it has been shown in [223] that adversarial attacks crafted to fool DL models trained on time-domain features do not necessarily transfer to DL models trained using frequency-domain features. By leveraging differences in channels, it is also viable to seek the objective of correct signal classification at one receiver while fooling the signal classifier at another receiver [28], [31], [176], [182], [224].
- *Exploitation and mitigation of channel effects:* The adversarial perturbation may aim to fool a wireless signal classifier at one receiver while the perturbed signal can

be still decoded at the intended receiver with minimal loss of reliability [176]. This paradigm can be used to hide wireless signals (ranging from simple modulated signals to complex 5G signals) from eavesdroppers [182], [225]. While balancing the objectives of covertness and communication performance, the adversary may aim to preserve not only the bit error rate (BER) as in [31] but also the spectral shape of the perturbed signal as in [58]. In addition, forward error correction (FEC) - a building block in communication systems to detect and correct errors encountered due to noise, channel effects, and hardware impairments - may reduce the effect of adversarial attack while maintaining communication performance with the intended receiver [33]. On the other hand, the adversary may use multiple antennas to transmit perturbations to increase the effect of adversarial attacks [30].

- *Indirect influence on training and test data:* A wireless adversary cannot directly manipulate the training or testing data input to a classifier. It also cannot directly query a transmitter's classifier and obtain its classification results as opposed to the typical CV and NLP applications that often rely on API queries. In an over-the-air attack, the wireless adversary needs to monitor the actions in wireless communications and indirectly try to manipulate/influence the outcome of the DL model or initiate actions such as jamming the channel during sensing and data transmissions [34].
- *Heterogeneity in feature representation:* The coexistence of various communication systems introduces heterogeneity as a more diverse and complex feature representation in radio data [35]. This significantly affects the effectiveness of crafted perturbations. In this context, the adversary may also aim to fool the wireless signal classifier at multiple receivers (each possibly belonging to a different network) by transmitting a signal perturbation and relying on the broadcast nature of omnidirectional transmissions to reach and jointly affect multiple classifiers. Then, the perturbation needs to be determined by accounting for multiple channels to different receivers [28]. In addition to dynamic channels, it is also possible that network traffic is of stochastic nature or traffic needs to be adapted to dynamic channel effects. Then, the generation of perturbations needs to be jointly considered with queue stability with the extended goal for the adversary to reduce the stable throughput (the maximum achievable throughput while keeping the packet queues stable [226]).

IV. Categorization of Adversarial Machine Learning Attacks on Wireless Communications

We present a categorization of AML attacks to aid the understanding of key concepts of AML and provide an overview of AML attacks from various perspectives borrowed from [29], [35], [51], [52], [53], [55], [231] for wireless communications and [23], [65], [67], [68], [69], [85], [232] for other domains, especially CV. In particular, we discuss

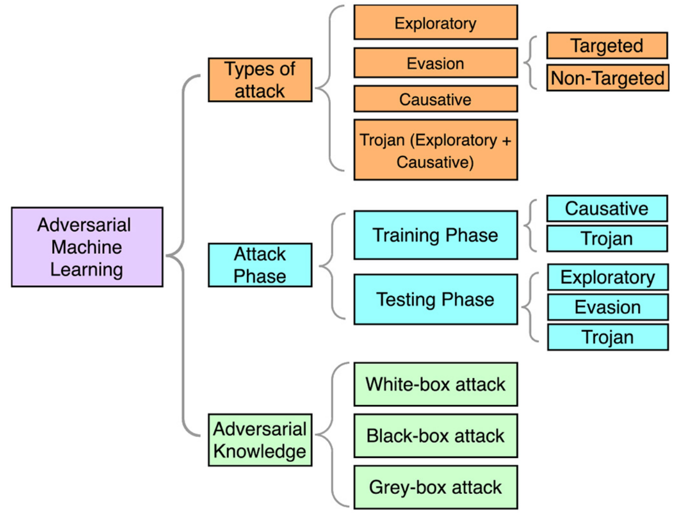

Fig. 5. Categorization of attacks based on adversarial machine learning.

adversarial attacks in wireless communications under three main categories (see Fig. 5) that are defined according to the influence - the type of attack, the attack phase, and the amount of knowledge the adversary has about the victim model. Table IV lists previous works of AML attacks on wireless communications under these categories.

A. Categorization Based on Influence—Types of Attack

- *Exploratory attacks:* Also called inference attacks, exploratory attacks seek to fathom the inner workings of ML algorithms/models by collecting training data and imitating the ML model functionally with similar types of inputs and outputs, namely building a surrogate (shadow) model [233], [234]. The exploratory attack is typically the leading step prior to subsequent attacks as it aims to "explore" the victim model using techniques such as active learning [235], [236] or augment the limited information with generative adversarial networks (GANs) [237], [238], [239], [240]. Adversarial attacks designed with the surrogate model are known to transfer to the target model (known as the transferability property [241]). In a wireless communications scenario, the adversary can learn the transmission patterns of the victim communication system by observing the spectrum over the air. The effect of wireless channels on surrogate models was studied in [59].
- *Evasion attack:* Also called adversarial attacks, evasion attacks aim to fool ML models into making wrong decisions by manipulating the input test data [14], [17]. Evasion attacks have been applied to wireless communications in terms of fooling classifiers used for spectrum sensing [34], modulation recognition [132], autoencoder-based end-to-end communication systems [54], channel state information (CSI) feedback for massive MIMO [89], [242], channel estimation [243], [244], and initial access in directional communications [245]. One challenge in the wireless domain is that the adversarial

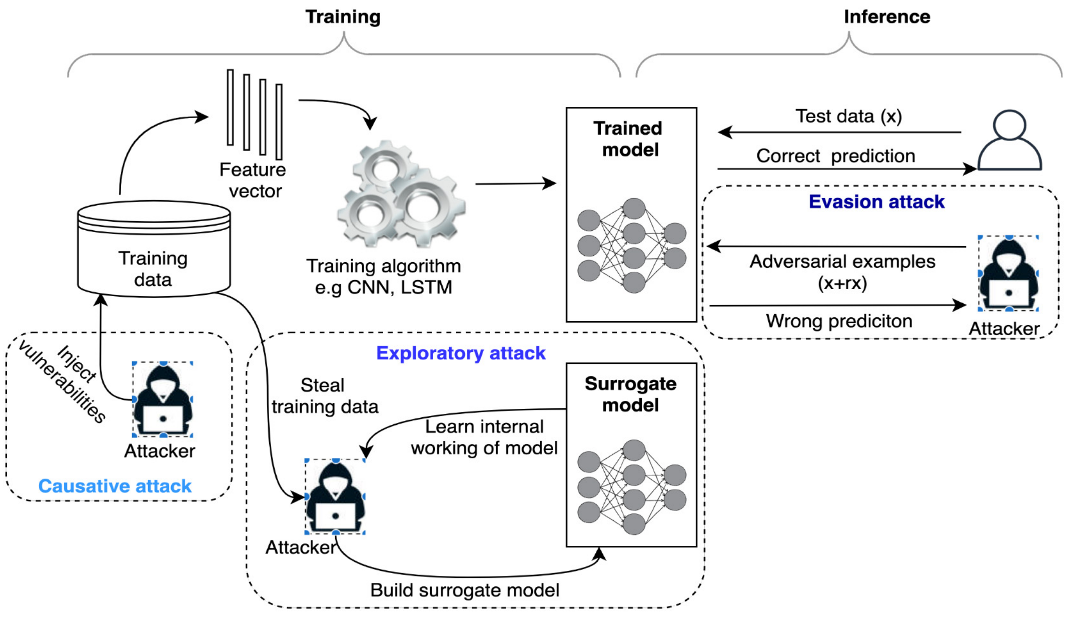

Fig. 6. Attacks in adversarial machine learning [250].

perturbations observed by the target classifier are received subject to channel effects such that these perturbations need to be crafted by accounting for channel effects [28], [31], [32], [56]. These attacks are typically stealthier and more energy-efficient than traditional jamming attacks (such as the one that solely aim to cause interference to data transmission, e.g., [76], [77], [246]) as they only need to transmit low-power signals over a short period of time to confuse the ML algorithms in their decision making.

- *Causative attack:* Also called poisoning attacks, causative attacks aim to manipulate the training process of ML models by injecting vulnerabilities such as false training data to the ML models [247], [248]. Causative attacks are effective against DNNs as they are highly susceptible to inaccuracies in training data. An example of a causative attack in wireless communications is spectrum poisoning where the adversary exploits the (re)training process of the ML classifier so that the ML model is poorly (re)trained. Examples of causative attacks in wireless communications include attacks on spectrum sensing [34], cooperative spectrum sensing [228], and IoT systems [229], [249].
- *Trojan attacks:* Also called backdoor attacks are combinations of evasion and causative attacks, where the adversary injects triggers (backdoors) to training data and then activates them for some input samples in test time [251], [252]. Trojan attack has been considered against wireless signal classifiers in [173] by formulating controlled phase shifts as backdoors and adding them into the transmitted and received RF data samples.

Figure 6 gives a pictorial representation of AML attack models and shows how the different attacks work, as well as the general procedure for attack formation.

B. Categorization Based on Attack Phase

- *Training phase:* The causative and trojan attacks occur in the training phase of the model development. See Section IV-A for discussion on these attacks.
- *Testing phase:* The exploratory attack, evasion attack and trojan attack occur in the test time. See Section IV-A for discussion on these attacks.

C. Categorization Based on the Amount of Knowledge the Adversary Has About the Victim Model

- *White-box attack:* The adversary knows the training data, architecture, algorithm and/or optimization techniques such that it has full access to the trained model f and knows the input at the classifier. For an over-the-air white-box attack in wireless communications, the adversary also needs to know the channel between the adversary and the receiver, since the channel affects the perturbation perceived at the receiver [28], [31], [32], [59].
- *Black-box attack:* The black-box attack is a more realistic and more relaxed model for many security threats [253] as the adversary has neither knowledge about nor access to the training data and/or the trained model f. In that case, the adversary tries to deduce information from the returned results of the model. Black-box attacks typically use a surrogate (shadow) model, which is trained to perform an identical task as the target network during the inference attack. In wireless communications, the surrogate model strongly depends on the channel characteristics with respect to the adversary and the victim system [59].
- *Gray-box attack:* In a gray-box attack, the adversary has some knowledge of the data/algorithm as well as limited access to the model f. One example from the wireless

domain is not knowing the exact channel but knowing an estimate, the distribution or some statistics of the channel, e.g., Rayleigh vs. Rician fading, the value of pathloss exponent or the value of signal-to-noise-ratio (SNR), etc. [28].

D. Categorization Based on Type of Adversary's Goals

- *Targeted attack:* The adversary tries to produce the perturbation r_x that will cause specific errors in the output of a receiver by targeting a signal x to generate errors such that $f_w(x + r_x) = y'$ for a particular target class y'. For example, consider a 4-class modulation classification problem with BPSK, QPSK, QAM 16 and QAM 64 as classes and consider the adversary that seeks to fool the modulation classifier into classifying a signal modulated with QAM 16 as a QAM 64.
- *Non-targeted attack:* The adversary tries to produce the perturbation r_x that will cause errors in the output of the algorithm independent of the classes thereby reducing the confidence in the algorithm due to decreased accuracy. In our previous example of a 4-class modulation classification problem with BPSK, QPSK, QAM 16 and QAM 64 as classes, the adversary seeks to fool the modulation classifier into classifying a signal modulated with QAM 16 as any other available class such as BPSK, QPSK, or QAM 64, i.e., $f_w(x + r_x) \neq y$ for any $y = f(x)$.

E. Other Types of Categorization

Some other types of categorization for AML attacks include categorization based on the frequency of attack and categorization based on the degree of freedom the adversary has in accessing the system's input.

- In the attack frequency category [35], there is a one-step attack that computes the gradient of the loss function at a time to create perturbations. There is also an iterative attack that performs the same computation as a one-step attack multiple times to create the perturbations. This is enabled by perpetual access to the model and can be computationally intensive and costly.
- Based on the adversary's degree of freedom to the input of the system, there may be physical and digital attacks. In a physical attack, there is an indirect application of the input to the model which is different from the digital attack where the adversary can explicitly design the input of the model [54], [253]. We can also remove the requirement for the adversary to know the input test samples and craft input-agnostic adversarial attacks, also called Universal Adversarial Perturbation (UAP) [254]. UAP attacks can be built against wireless signal classifiers, where the adversary does not need to know about the transmitted signal [28], [132]. It is also possible to limit attacks to certain users, e.g., evasion attacks to fool classifiers only at a subset of receivers [28] and causative attacks launched from a selected subset of IoT devices to manipulate the training process at a fusion center [229].

These attack types may overlap in the description of an attack. For example, the authors in [29] studied both targeted, white-box evasion attack and non-targeted white-box evasion attacks for DL-based modulation classification.

V. Generation of Attacks in Wireless Communications

In this section, we describe the approaches to generate attacks in AML, including evasion attacks, exploratory attacks and causative attacks.

A. Generation of Evasion Attacks

Evasion attacks to generate adversarial examples are characterized as subtly crafted imperceptible perturbations r_x added to the original inputs x of the ML model such that the input to the classifier is $x' = x + r_x$ by solving the optimization problem (6) for r_x. In the wireless domain, the problem is changed by setting with $x' = h_{t,r}x + h_{t,a}r_x + n$, where x' is the received signal, x is the transmitted signal, r_x is the transmitted perturbation, $h_{t,r}$ is the channel gain from the transmitter to the receiver, $h_{t,a}$ is the channel gain from the transmitter to the adversary, and n is the receiver noise.

It is difficult to solve (6) due to the non-linearity of the DNN mapping. Instead, we can approximate the solution. This is achieved by maximizing the loss function $L(w, x, y)$ used in the training process of the classifier f_w (e.g., the cross-entropy loss function is $L(w, x, y) = -\log(1 + \exp(-f_w(x, y)))$ subject to the condition that the perturbation is upper bounded. Thus, the adversarial perturbation r_x can be found by solving

$$\begin{aligned} \max \quad & L(w, x + r_x, y) \\ s.t. \quad & \min ||r_x||_p \leq \epsilon, \end{aligned} \tag{7}$$

where ϵ is the upper bound on the adversarial perturbation. Various methods have been proposed to approximately solve (7), thus generating adversarial examples. These methods include but are not limited to one-step gradient-based method such as Fast Gradient Sign Method (FGSM) [17], and its iterative variants such as Basic Iterative Method (BIM) [253], Projected Gradient Descent (PGD) [190], Momentum Iterative Method [255], Box Constrained Limited-Memory BroydenFletcherGoldfarbShanno (L-BFGS) [14], and Carlini & Wagner (C&W). We describe some of these methods that have been used to generate adversarial attacks on wireless communications.

1) Fast Gradient Sign Method (FGSM) [17]: One popular method that has been used to generate adversarial examples in the wireless domain is the FGSM. The FGSM is a one-step gradient-based method developed on finding the scaled sign of the gradient of the cost function and aims at minimizing the strength of the perturbation. It is a computationally effective method for adversarial attack generation based on the postulation that the DNNs are prone to linear-type adversarial attacks because of their use of linear techniques for easier optimization. The FGSM-generated perturbation is achieved by linearizing the model's cost function $J(w, x, y)$ around the current value of x. FGSM generates an adversarial example that significantly resembles the original input x by starting with an original input to the ML model and adjusts it in the direction of the gradient of the loss function with respect to the input by an amount limited by a parameter ϵ stated as:

$$x^* = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y, w)), \tag{8}$$

where x^* is the crafted adversarial example, x is the original input, $J(x, y, w)$ is the cost function of the DNN with parameters w, y is the class associated with x, and ϵ is a small number to limit the perturbation [230]. Note that the Fast Gradient Method (FGM) is a generalization of FGSM to meet the L_2 norm bound $||x^* - x||_2 < \epsilon$ as

$$x^* = x + \epsilon \cdot \frac{\nabla_x J(x, y, w)}{||\nabla_x J(x, y, w)||_2}. \tag{9}$$

FGSM is also used in iterative methods such as the following.

- *Basic iterative method (BIM) [253]:* The BIM can be viewed as multiple steps of FGSM with a small step size as well as a clipping of results after each iteration to establish that the results are in an ϵ-neighborhood of the original input data. The iteration for the BIM is given by

$$x_i^* = \text{clip}_{x,\epsilon}\left(x_{i-1}^* + \epsilon \cdot \text{sign}\left(\nabla_x J\left(x_{i-1}^*, y, w\right)\right)\right), \tag{10}$$

  where $\text{clip}_{x,\epsilon}$ clips the values of the adversarial sample to an ϵ-neighborhood of the original sample x.
- *Projected Gradient Descent (PGD) [190]:* PGD is a multi-step variant of FGSM on the negative loss function that helps initiate adversarial attacks with the aim of examining or analyzing the performance of a neural network from an optimization point of view. It puts a limit on the total perturbation in L_∞-norm from $x_0^* = x + U_n(-\epsilon, \epsilon)$, where $U_n(-\epsilon, \epsilon)$ is the uniform random perturbation. In t iterations, x is determined as

$$x_t^* = \Pi_{B_\epsilon(x)}\left(x_{t-1}^* + \epsilon \cdot \text{sign}\left(\nabla_x J\left(x_{t-1}^*, y, w\right)\right)\right), \tag{11}$$

  where $\Pi_{B_\epsilon(x)}$ is the projection operator. The procedure of adversarial example generation includes adding noise, computing gradient, stepping, and projecting back.
- *Momentum Iterative Method (MIM) [255]:* The MIM was developed to solve the problems related to over-fitting and local minima. It helps speed up the gradient descent iterations. In scenarios where the performance of FGSM is inhibited due to the presence of strong interference and distortion, the MIM applies momentum to the iterative attacks so as to stabilize and maintain the correct direction of the update with generalization of adversarial examples.

2) L-BFGS Method [14]: Box-constrained L-BFGS is an example of optimization-based method for generating adversarial examples. For an instance of x, L-BFGS finds a different x^* that is very much alike to x based on the L_2 distance but is labeled differently by the classifier. This problem is laborious and is modeled as the following constrained minimization:

$$\begin{aligned} \min \quad & ||x - x^*||_2^2, \\ s.t. \quad & f_w(x + r_x) = y', \\ & (x + r_x) \in [0, 1]^n, \end{aligned} \tag{12}$$

where y' is the target class. To solve (12), it is transformed to

$$\begin{aligned} \min \quad & \lambda \cdot ||x - x^*||_2^2 + J\left(x, y', w\right) \\ s.t. \quad & x^* \in [0, 1]^n, \end{aligned} \tag{13}$$

where $J(x, y', w)$ is a function mapping an instance of x to a positive real number using the loss function. In (13), the cross-entropy loss function is typically used as the loss function J and line search is used to find the positive constant λ that generates an adversarial example of minimum distance.

B. Generation of Exploratory Attack

In an exploratory attack, an adversary seeks to understand the internal operations on an ML model and leverage on such knowledge to attack the ML model and ultimately undermine the effectiveness of the ML-based system. This is achieved by developing a surrogate (or shadow) model that can mimic the operation of the original ML model. The approach is to steal training data and build models that have the same (or similar) output as the original model. With the information on how the original model functions, the attacker can make informed decisions as to how to attack and alter the overall performance of the ML-based system. For example, in [227], an exploratory attack is launched on the cognitive radio as a preliminary step to learn when and how to jam the cognitive radio transmissions. The adversary senses a channel, captures the transmitters decisions and learns the pattern of successful transmissions by tracking acknowledgments from the receiver, and launches an exploratory attack to build a DL classifier (surrogate model) that is functionally equivalent to the one at that transmitter. In another example [51], an adversary observes the spectrum and develops a DNN to infer the channel access algorithm used by an IoT transmitter and predict the outcome of the transmissions before jamming it. Exploratory attacks are also launched to infer reinforcement learning mechanisms used in wireless channel access and jam the underlying communications. [60], [61] designed a DRL-based jamming attacker that employs a dynamic policy and aims to minimize the channel access accuracy of a DRL-based dynamic channel access user. The DRL attacker is able to observe the victim's interaction with its environment for a period of time and learns the activity pattern. The attacker has no prior information about the channel switching pattern on a victim's action policy and both DRL-based systems can interact with each other, retrain their models and adapt to the opponent's policy. Also, [62], [63] introduced a reinforcement learning (RL) algorithm to disrupt 5G radio access network slicing (based on another RL algorithm [256], [257]) by observing the spectrum and developing an RL-based surrogate model. This model decides which resource blocks to jam in order to attain a high number of failed network slicing requests. Jamming a resource block causes a reduction in the RL algorithm's reward thus, affecting the model's performance as the reward is used to update the RL algorithm.

Although the exploratory attacks described above follow the same paradigm, there are no standard attack generation methods as we have in evasion attacks. All the methods use a common approach based on learning the internal workings of

TABLE V
REVIEW OF PREVIOUS WORKS ON AML ATTACKS IN MODULATION CLASSIFICATION

Paper	Attack Type	Objective	Attack Method	DL Problem	Data used
[29]	White-box, Black-box, Targeted, Untargeted, Evasion	Explore how to launch a realistic evasion attack by taking into consideration the channel effect as well as the power constraint at the adversary.	FGM, UAP	Modulation classification	RML2016.10A [260]
[175]	Evasion	Test the attack efficiency of two types of adversarial attacks.	FGSM, Box-constrained L-BFGS	Modulation classification	RML2016.04C [4]
[35]	Evasion, Black-box, White-box	Explore the feasibility and effectiveness of adversarial attacks and determine the optimal perturbation level for attack invisibility and success.	FGSM, PGD, BIM, MIM	Modulation classification	RML2016.10A [260]
[173]	Trojan attack	Introduce a new attack that slightly manipulates training data by inserting Trojans (i.e., triggers) to a few training data samples and then activate them later in test time.	Embed Trojans in training data and trigger it in test time	Modulation classification	RML2016.10A [260]
[55]	Evasion	Evaluate the vulnerabilities of raw I/Q based modulation classification.	FGSM	Modulation classification	RML2016.10A [260], three synthetic datasets using GNU radio
[31], [176], [261]	Evasion	Minimize the accuracy of the intruder in determining the modulation scheme used by a transmitter by perturbing the channel input symbols.	PGD	Modulation classification	Locally generated data
[172]	Gray-box	Mitigate adversarial examples in RF classifiers using autoencoder preprocessing.	FGSM	Modulation classification	RML2018A [108]
[30]	Evasion White-box Targeted	Investigate the use of multiple antennas to generate multiple concurrent perturbations over different channel effects (subject to total power budget) to the input of DNN-based modulation classifier at a wireless receiver.	FGM	Modulation classification	RML2016.10A [260]
[132]	Evasion, White-box, Black-box	Show the susceptibility of DL models to adversarial attacks and present practical methods for crafting adversarial examples in modulation classification.	PCA-based UAP, FGM-modified bisection search	Modulation classification	RML2016.10A [260]
[182]	Evasion, Targeted, White-box	Fool a DL-based eavesdropper by a cooperative jammer transmitting adversarial perturbation so as to hide 5G communications from the eavesdropper.	FGM	Modulation classification and 5G system	Locally generated data using MATLAB 5G toolbox
[33]	Evasion	Extend the use of communication-aware evasion attack through the use of forward error correction (FEC).	Perturbation created by Adversarial Mutation Network	Modulation classification	Synthetic data generated using Liquid DSP
[174]	Evasion, Untargeted, White-box, Targeted	Examine how the DL-based automatic modulation classifier breaks down in the presence of an adversary with direct access to its inputs.	FGSM, MI-FGSM	Modulation classification	RML2016.10A [260].
[28]	Evasion, Black-box, Untargeted, White-box, Targeted	Explore how to design realistic adversarial attacks in the presence of realistic channel effects and multiple classifiers at different receivers.	FGM, UAP	Modulation classification	RML2016.10A [260]
[32]	White-box, Untargeted Targeted	Provide a generalized wireless AML problem (GWAP) and experimental evaluation of AML attacks launched against wireless DL systems	AML waveform jamming, AML waveform synthesis	Modulation classification, Radio Fingerprinting	RML2018.01A [108], 1000-device dataset of WiFi and ADS-B transmissions
[58]	Evasion	Introduce spectral deception loss metric implemented during training to make the spectral shape more in-line with the original signal.	Adversarial Mutation Network (AMN)	Modulation classification	Locally generated data
[176]	Evasion	Minimize the accuracy of an intruder while still ensuring the successful decoding of a signal by the intended receiver.	PGD, uniform random noise of L_2-norm to a block of 128 I/Q symbols	Defense against modulation detection	Locally generated data
[262]	Evasion	Study the effect of adversarial attacks on ML-based AMC model using different data-driven subsampling strategies	Carlini-Wagner	Modulation classification	RML2016.10B [260]
[263]	white-box, black-box	Present an input-agnostic adversarial attack that is undetectable and robust to removal	UAP	Modulation classification	RML2016.10A [260]

the original model, building a surrogate model and attacking the ML-based model using the knowledge from the surrogate model. Table IV shows examples of previous works on exploratory attacks.

C. Generation of Causative Attack

The causative attack seeks to downgrade the ability of a ML system to perform optimally by injecting vulnerabilities into the training process. These injected vulnerabilities can be as a result of the adversary by intentionally meddling with the label of the data or by crafting and embedding unique learnable features. The model learns these unique features which causes it to deviate from its original objective, thus, causing considerable degradation in the ability of the model to make good predictions. The injection of vulnerabilities might be during the initial model development or the model re-training process. Since the attack occurs before model training, the contamination affects any type of ML model and efforts on parameter tuning yields little or no improvement on the model's ability to make good predictions [250]. As we have observed in the exploratory attack, there are no standard attack generation methods but the paradigm of poisoning the training data is consistent in all forms of causative attacks. Table IV shows some previous work on causative attack.

VI. APPLICATION AREAS OF ADVERSARIAL MACHINE LEARNING IN WIRELESS COMMUNICATIONS

Grouped according to the type of DL problem, we discuss and review various studies on AML in wireless communication and present them in Tables V (for modulation classification), Table VI (for spectrum sensing), Table VIII (for signal

TABLE VI
REVIEW OF PREVIOUS WORKS ON AML ATTACKS IN SPECTRUM SENSING

Paper	Attack Type	Objective	Attack Method	DL Problem	Data used
[52], [228]	Exploratory	Launch an exploratory attack on cognitive radio as a preliminary step before jamming.	Jam legitimate user transmissions	Cognitive radio channel access	Locally generated data
[53]	Spectrum poisoning, Exploratory	Launch an over-the-air attack to infer the transmitter's behavior and falsify the spectrum sensing data by transmitting perturbations over the air.	Inject perturbation to the channel	Spectrum sensing	Locally generated data
[51]	Exploratory, Evasion, Causative, Black box	Build a DNN classifier to infer the channel access algorithm of a transmitter and predict the transmission result, and develop a defense mechanism to confuse the adversary in its attack strategy.	Jamming, Spectrum poisoning, Priority violation	Spectrum sensing	Locally generated data
[229]	Black-box, Exploratory	Propose Learning-Evaluation-Beating(LEB) attack against cooperative spectrum sensing (extension of spectrum sensing data falsification attack on cognitive radio).	LEB attack	Cooperative spectrum sensing	Dataset collected from 5282 locations [264], Data from 5 locations using USRP N210
[61]	Black box Gray box	Develop deep reinforcement learning (DRL)-based dynamic channel access system and DRL-based jamming attack to study the sensitivity of DRL-based wireless communications to adversarial attacks.	DRL-based attacker use dynamic policy to minimize the victim's channel access	Dynamic spectrum access (DSA)	Locally generated data.
[34]	Exploratory, Evasion, Causative	Launch over-the-air spectrum poisoning attacks by learning first the transmitter's behavior.	Attacker jams the transmitter's sensing period	Spectrum sensing, DSA	Locally generated data
[60]	Exploratory	Compare the performances of two adversarial policies: feed-forward neural network (FNN) and DRL policies on the performance of a DRL-based dynamic channel access agent.	Jamming of a victim model	Dynamic spectrum access	Locally generated data
[265]	Exploratory, Evasion	Present AML attack on spectrum sharing of 5G communications with incumbent users.	Manipulate inputs to the classifier by transmitting during spectrum sensing	Spectrum sharing	Locally generated data

TABLE VII
REVIEW OF PREVIOUS WORKS ON AML ATTACKS RESOURCE ALLOCATION AND MANAGEMENT

Paper	Attack Type	Objective	Attack Method	DL Problem	Data used
[179]	Black-box	Develop defense strategies against DRL-based jamming attacks on a DRL-based dynamic multichannel access agent.	DRL-based jamming attacker	Multichannel access	Locally generated data
[246]	non-targeted	Investigate the vulnerability of a DNN used for mmWave beam prediction as part of the initial access process in 5G and beyond communications.	FGM, adversary tries to change the beam to one of the worst beams	mmWave beam prediction in 5G	Locally generated data
[266]	Exploratory	Study the vulnerabilities of DRL agents that perform both dynamic channel access and power control in wireless interference channels to adversarial attacks.	DRL-based adversarial jamming attack	dynamic channel access and power control	Locally generated data
[243]	white-box, black-box	Show that adversarial attacks can break DL-based power allocation in the downlink of a massive multiple-input-multiple-output (maMIMO) network.	(FGSM), momentum iterative FGSM, and PGD	Power allocation in a maMIMO	DL power allocation in massive MIMO dataset [267]
[60]	Exploratory	Investigate the learning-based wireless jamming attacks on deep reinforcement learning policies on dynamic multichannel access.	Feed-forward neural network and DRL based jamming attacks	Dynamic channel access	Locally generated data
[62]	Exploratory	Present an over-the-air attack to manipulate the RL algorithm and disrupt 5G network slicing.	Adversary builds an RL-based model and jams selected resource blocks	Resource allocation for 5G network slicing	Locally generated data
[63]	Exploratory	Present flooding attack on 5G network slicing.	Adversary crafts fake network slicing requests to consume the 5G radio access network resources	Resource allocation for 5G network slicing	Locally generated data
[268]	Causative, Exploratory, targeted	Check the effect of AML based attacks on power allocation where the base station (BS) allocate its transmit power to multiple orthogonal subcarriers by using a deep neural network (DNN) to serve multiple user equipments (UEs).	Simplified Analytical Gradient-Based Attack and DNN Gradient-Based Attack to change the DNNs input	Power allocation	Locally generated data

classification), and Table IX (for other areas in the wireless communication domain).

A. Modulation Classification and Signal Classification

Modulation classification – the process between the detection and demodulation of a signal – is an important step towards developing an intelligent radio receiver and has been extensively studied [258]. Recently, the research in this area has gained new attention with the use of DL. Reference [4] showed that DL algorithms for modulation classification using RF data outperform the use of expert features and higher order moments. AML for modulation recognition has been extensively using evasion and/or trojan attacks. Reference [35], [132], [175] showed the susceptibility of DL models to adversarial attacks on modulation classification under different scenarios such as data with noise and variation in SNR level. Reference [55], [132] pointed out that the adversarial examples

TABLE VIII
REVIEW OF PREVIOUS WORKS ON AML ATTACKS ON SIGNAL CLASSIFICATION

Paper	Attack Type	Objective	Attack Method	DL Problem	Data used
[180]	Black-box, Exploratory	Show that DL-based signal classifiers are vulnerable to privacy threats due to over-the-air information leakage of their model.	Membership inference attack (MIA)	Wireless signal classification	Locally generated data
[59]	Exploratory, White-box, Targeted	Investigate the channel effects on surrogate model built by an adversary that uses the over-the-air observed spectrum data.	Maximum Received Perturbation Power (MRPP) attack [29]	Wireless signal classification	RML2016.10A [260] & locally generated data
[231]	Evasion	Develop methods for detection of adversarial perturbations used against wireless signal classifiers.	FGSM	Survey application & spectrum sensing	RML2018A [108], over-the-air Bluetooth, WiFi, ZigBee datasets
[178]	Evasion, Untargeted, White-box	Fool a DL-based signal classifier (authenticator).	Reinforcement learning-based attack	Signal authentication	Simulated data and testbed data from Pluto SDRs
[57]	Targeted, Black box	Analyze the effect of adversarial examples against Deep RF classifiers and also develop several defense mechanisms.	Carlini-Wagner attack	Signal & , protocol classification	Locally generated WiFi, Bluetooth & ZigBee data. RML2018A [108]
[263]	white-box, black-box	Present an input-agnostic adversarial attack that is undetectable and robust to removal.	UAP	Signal Detection in OFDM Systems	RML2016.10A [260]

TABLE IX
REVIEW OF OTHER PREVIOUS WORKS ON AML ATTACKS IN WIRELESS COMMUNICATIONS

Paper	Attack Type	Objective	Attack Method	DL Problem	Data used
[54]	Black-box, Exploratory, Evasion	Show that end-to-end learning of communication systems through autoencoder can be extremely vulnerable to physical adversarial attacks.	Iterative method based on UAP	End-to-end autoencoder communication systems	Locally generated data
[230]	Exploratory	Employ AML techniques to launch attacks against the IoT data fusion process.	Partial-model attack	IoT data fusion/aggregation process	Synthetic data from Gaussian distributions
[89]	White-box, Untargeted	Present a method to craft adversarial attack and show its effect on DL-based channel state information (CSI) feedback process.	Attacker modeled by a bias layer between the encoder and decoder of CsiNet	Massive MIMO CSI feedback process	Locally generated data generated using COST 2100 channel model [274]
[265]	Exploratory, Spoofing	Present adversarial attack on signal authentication in network slicing.	Spoofing attack on the network slicing application	Physical layer authentication of 5G device	Locally generated data

for modulation classification are more effective than perturbations crafted by adding Gaussian noise. Table V shows AML studies focused on modulation classification. Similar to the modulation classification problem, AML has also been applied to wireless signal classification. Signal classification aims to identify the transmitting signal in a spectrum. Table VIII highlights relevant work on AML for signal classification.

B. Spectrum Sensing

Most of the AML studies focused on spectrum sensing and dynamic spectrum access (DSA) use exploratory attacks in which the attacker learns the operational pattern of the DNN model and builds a surrogate model which gives the same or similar output as the original classifier when given the same input. Spectrum sensing and DSA are key components of cognitive radio systems for efficient discovery and use of spectrum to achieve situational awareness [53]. Given the broadcast and over-the-air nature of wireless communications, there are various attacks that attempt to undermine the spectrum sensing results. This can include causative attack to change the sensed features as well as priority violation attacks to violate priorities in channel access by pretending to have higher priority and transmitting during the sensing phase [51]. AML attacks on spectrum sensing typically involve an adversary that develops a surrogate model and can correctly predict the outcome of the original DL model. Such an adversary jams the successful transmissions, thus reducing the success rate of the original model [52], [60], [227]. As shown in [34], [52], [227], [228], these attacks are more energy-efficient and more effective than random jamming. Table VI highlight the studies on AML for spectrum sensing and DSA applications.

C. Resource Allocation and Management

The need to satisfy the demand for faster response time, increased reliability of wireless transmissions as well as larger bandwidth has made the use of efficient algorithm and approach for resource allocation and control such as power control, and beamformer design increasingly important. ML/DL techniques applied to achieve optimal resource usage are also susceptible to adversarial attacks. Table VII shows some studies on AML for resource allocation and management.

D. Other Areas in Wireless Communications

AML attacks have been also applied to other areas in wireless communications. This includes attacks on IoT data fusion process [229], IoT device authentication [268], physical-layer signal authentication [269], [270], spectrum monitoring [271] with federated learning [272]. A review of the AML studies on other areas in wireless communication is presented in Table IX.

VII. Detection and Mitigation of AML Attacks in Wireless Communications

Next, we present various methods that have been used to defend against AML attacks on wireless communication systems.

A. Train With Adversarial Perturbations

Training the DNNs with adversarial examples is a popular method of mitigating the effect of adversarial attacks and ensuring robustness [190]. Adversarial training consists of generating adversarial examples according to one or more attack methods and retraining the DNNs with labeled adversarial examples. The goal is to prevent an adversary from identifying the structure of the signals transmitted so as to prevent it from disrupting the communication system. This defense mechanism has been applied to protect wireless signal classifiers against evasion attacks [28], where randomized smoothing [274] used in training time increases the robustness of classifiers against adversarial attacks later in test time.

There are some challenges that pertain to adversarial training; some of these include a scenario where the adversary uses a different attack from the one used in training. In addition to that, an adversary can create new perturbations using a model trained with adversarial examples. However, training with adversarial examples typically reduces the performance of DL models on unperturbed signals. In trying to overcome these challenges, training can be performed with multiple adversarial examples [190]. In addition, certified defense [275] can be built in train and test times to ensure statistical significance of classification results in the presence of adversarial attacks in test time. Certified defense has been used in [28] to provide performance guarantees for wireless signal classifiers in the presence of adversarial examples. Another way to mitigate adversarial attacks during training is pre-training of the DL based wireless signal classifiers by an autoencoder such that the trained model becomes robust to the deceiving effect of adversarial examples [172].

B. Statistical Approaches

- *Peak-to-average power ratio (PAPR) of RF samples:* One way of detecting adversarial examples it to leverage the properties of radio frequency (RF) data and apply a statistical test by taking advantage of the PAPR of digitized RF samples of the received signals [230]. The PAPR distribution is used in the wireless domain to depict a specific class of signal's footprint such that if the inferred output of an input signal is categorized to a particular class and its PAPR statistic contradicts with high level of confidence, further analysis and assessment should be carried out to reach a true conclusion. To determine if an adversarial example is present in a wireless communication environment, the Kolmogorov-Smirnov (KS) two-sample test (performed by collecting a sample sized input and then computing and evaluating the PAPR distributions for the samples) can be used to ascertain if the PAPR is similar to the statistic of an adversarial example or a legitimate example [230].
- *Use the softmax outputs of the DNN classifier:* This is also a KS statistical test for adversarial example detection. The softmax output of the DNN classifier is used to determine if adversarial examples have brought about a change in distribution using the statistics of the final layer of a neural network. Typically, the statistical size of the input data is a good determination of the quality of the statistical test. The effectiveness of the method depends on the type of waveform and propagation channel, as discussed in [230].
- *Statistical methods to detect adversarial triggers:* Adversarial triggers such as Trojans (or backdoors) inserted to training data can be detected by statistical methods such as clustering and median absolute deviation (MAD) algorithm. In particular, the MAD algorithm computes the median of the absolute deviations from the data's median, namely $\mathrm{median}(|x_i - \hat{X}|)$, where $\hat{X} = \mathrm{median}(X)$, to discover outliers, and is shown in [173] to be effective against Trojan attacks on wireless signal classifiers.

C. Randomizing the DL Algorithm by Adding Small Variations to Classifier Outputs

This approach adds slight variations to the output of the DNN classifier and deliberately performs some (typically a small number of) erroneous actions so as to deceive the adversary and defend against exploratory attacks. For example, the transmitter can intentionally utilize a busy channel for transmitting or skip a transmission opportunity in an idle channel such that the adversary that observes the spectrum cannot train a reliable surrogate model. This defense mechanism has been applied to protect wireless signal classifiers against inference attacks (before launching jamming attacks) in [227] as well as against evasion and causative attacks in [34]. Note that network uncertainty (e.g., due to dynamic channel and traffic effects) and randomization or obfuscation-based defense methods have also been leveraged before against traditional network attacks such as jamming and traffic inference to confuse the adversaries [76], [77], [246], [276]. Defense against AML attacks on reinforcement learning follows a similar diversification approach to protect wireless communications [61], [179]. The defender may alternate among different channel access decisions (derived through proportional-integral-derivative (PID) controller and imitation learning) to increase the uncertainty at the adversary. Another defense approach is to adopt orthogonal policies in reinforcement learning to prevent the adversary from quickly switching between its imitated policies. On top of these mitigation policies, it is also viable to detect the adversaries by monitoring changes in rewards and distinguishing the spectrum environment changes from adversarial jamming attacks [61], [179]. Table X reviews the defense strategies that have been used by previous works in the wireless domain.

TABLE X
SUMMARY OF DEFENSE STRATEGIES AGAINST AML ATTACKS IN WIRELESS COMMUNICATIONS

Paper	Defense Strategy
[34], [51], [52], [176], [228]	Deliberately makes a small number of incorrect transmit actions so as to fool the adversary and prevent it from building a reliable classifier, e.g., transmit in a busy channel.
[31], [176]	Slightly modifies the transmitted signal; enough to fool an intruder and allow for correction using the error correction modules of the receiver.
[172]	Uses an autoencoder for the DL classifier to remove non-salient features thereby removing perturbations.
[173]	Uses the MAD algorithm and two-step technique of dimensionality reduction and clustering via t-distributed stochastic neighbor embedding (t-SNE) to detect Trojan triggers.
[57], [177], [231]	Applies KS Statistics that uses the peak-to-average power ratio or the softmax output of the ML model for distribution shift detection.
[28]	Applies randomized smoothing in training time to increase the robustness of classifiers against adversarial attacks in test time and ensures statistical significance of classification results in the presence of adversarial attacks in test time.
[61], [179]	Applies a diversification approach by alternating the reinforcement learning algorithm among different channel access decisions (via PID controller and imitation learning) and adopting orthogonal policies to restrict the adversary's response.
[176]	Adding a constrained perturbation to the signal and ensuring that the signal remains decodable by the oblivious legitimate receiver and minimizing detection accuracy at the intruder.
[57]	Pre-training the DL classifier using an autoencoder which is expected to filter out non-salient features that may slightly lower the accuracy of classification and makes the classifier more robust to adversarial or environmental corruption.
[229]	Introduced influence-limiting defense by using decision-flipping influence which indicates the probability of finding a malicious input that modifies the decision output given the unperturbed data by changing part of the unperturbed signal.

VIII. FUTURE OUTLOOK OF ADVERSARIAL MACHINE LEARNING IN WIRELESS COMMUNICATIONS

As ML/DL becomes the core of current and emerging communications systems such as 5G and 6G, ML/DL itself becomes susceptible to adversarial effects. To build on the promise of intelligent operations and efficient resource management with the application of ML in dynamic wireless environments, it becomes imperative to develop ML models that are secure, resilient and robust to attacks by taking into account the unique properties of wireless communications. An AML attack model that is carefully characterized by considering wireless properties is expected to be the foundation of ML driven wireless security research and development for the existing and emerging communication systems. In this section, we discuss some important ideas that can help in realizing the promise of robust models that are effective even in the presence of adversaries.

- *Wireless communication dataset:* The limitation of standardized real-world datasets that adequately represent real life scenarios in the wireless communication domain is a challenge that needs to be solved. Compared to other domains such as CV and NLP, there are just a handful of publicly available ML/DL datasets in the wireless domain to work with such as [4], [108], [277], [278], [279], [280]. Typically, these datasets do not include adversarial effects recorded. It is important for the research community to develop more publicly available datasets, representing different scenarios including not only variations in channel, interference type and waveform, but also AML attacks. This approach will help with benchmarking solutions for robust development and evaluation of the DL models for wireless versions of AML attacks. From literature reviewed, a large ratio of the work on AML in wireless communication is in the area of modulation classification. This is largely because modulation classification datasets (e.g., [4], [108]) are the seemingly standardized datasets for Ml/DL studies in the wireless communications domain. Although more recent datasets are available such as [277], [278], [279], [280] as well as testbed efforts such as [281] expand datasets, it is critical to implement the AML attacks and corresponding defenses with embedded radio platforms to assess the real channel and radio hardware effects.
- *Robust features:* The need for robust features in developing ML/DL models in wireless communication is also an important consideration for future systems. Reference [282] laid the analytical foundation and showed via experimental data that adversarial attacks are very effective because most existing ML/DL models are not developed using robust features. Although the discussion of feature engineering and identifying robust features are on-going especially in CV and NLP domains, new techniques must be developed to identify the most important features which are robust to the effect of adversarial attacks in wireless communication systems. An important research direction for AML in wireless communications is to identify robust features in datasets, disentangle non-robust features from robust features, train models that generalize well, and are robust to adversarial attacks in wireless communications.
- *Certifiable defense:* The design and development of certifiable defense mechanisms for AML is another important component of the work that needs to be achieved to realize the full potential of ML/DL in wireless communications domain. Various defense mechanisms have been proposed, including adversarial training and use of statistical methods but most of these defenses become ineffective with more powerful adversaries. This has led to the need for certifiable defense mechanisms, where prediction on the test data can be verified to be a constant within a small region around the input data. Although there are some efforts that present certifiable defense using randomized smoothing [28] with reasonable performance, more rigorous evaluation of such approaches as well as development of other methods are needed to provide security guarantees for ML models used in wireless communications.

Although important areas of development to ensure good performance from robust ML/DL models have been discussed in this section, for the safe adoption of ML/DL-based models into future communication systems, there is also a need to move from black-box models to interpretable models. Such paradigm as seen today includes physics-guided ML [283], where the fundamental laws of physics is infused into the learning process to improve generalizability and explainability, and ensure robustness to attacks, as needed in wireless communication systems.

IX. Conclusion

The recent trends observed in this survey show that the AML attacks can effectively disrupt wireless communications. Compared to the traditional jamming attacks or the addition of noise/interference, adversarial attacks are more difficult to detect operating with small spectrum footprint, more energy efficient, and overall more effective. Furthermore, as wireless communications evolves into smarter systems as in 5G and 6G, the AML attacks also become smarter. The attacks have evolved from adversaries that send jamming signals randomly to overwhelm a communication system to those that develop surrogate models to know the internal workings of the original model and send out small perturbation signals to fool the ML/DL models and disrupt wireless communications.

The goal of AML is to support safe adoption of ML/DL solutions to the emerging applications in the presence of adversaries. Wireless communications strongly benefit from ML/DL applications in terms of learning from spectrum data and solving complex tasks. Various studies reviewed in this paper have established that AML attacks on the ML/DL-driven wireless systems are effective and pose serious threats leading to major performance degradation. This review attempts to present the current state of the art in the area of AML attacks specific to the wireless communications domain and help the research community identify the research accomplishments as we forge ahead to investigate methods by which efficient data driven intelligent systems can be built while taking into account the unique nature of wireless communications, robustness, resilience, and security of ML/DL based methods. We summarized the latest research efforts by reviewing the attack types, adversarial example generation, and defense strategies, established a taxonomy across different attack categories, and discussed the future outlook to AML in wireless communications. This is an emerging research area and there is an urgent need for further investigations to quantify the impact of AML attacks and detect and mitigate them to enable the safe adoption of the promising ML/DL solutions for wireless communications.

Acknowledgment

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. National Science Foundation (NSF) or the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) or the U.S. Government.

References

[1] T. Erpek, T. J. O'Shea, Y. E. Sagduyu, Y. Shi, and T. C. Clancy, "Deep learning for wireless communications," in *Development and Analysis of Deep Learning Architectures*. Cham, Switzerland: Springer, 2020, pp. 223–266.

[2] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 648–664, Dec. 2018.

[3] L. J. Wong, W. H. Clark, B. Flowers, R. M. Buehrer, A. J. Michaels, and W. C. Headley, "The RFML ecosystem: A look at the unique challenges of applying deep learning to radio frequency applications," 2020, *arXiv:2010.00432*.

[4] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks*, C. Jayne and L. Iliadis, Eds. Cham, Switzerland: Springer Int. Publ., 2016, pp. 213–226.

[5] F. Liang, C. Shen, and F. Wu, "An iterative BP-CNN architecture for channel decoding," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 144–159, Feb. 2018.

[6] H. Jafari, O. Omotere, D. Adesina, H. Wu, and L. Qian, "IoT devices fingerprinting using deep learning," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2018, pp. 1–9.

[7] T. Jian *et al.*, "Deep learning for RF fingerprinting: A massive experimental study," *IEEE Internet Things Mag.*, vol. 3, no. 1, pp. 50–57, Mar. 2020.

[8] K. Merchant, S. Revay, G. Stantchev, and B. Nousain, "Deep learning for RF device fingerprinting in cognitive communication networks," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 160–167, Feb. 2018.

[9] K. Davaslioglu, S. Soltani, T. Erpek, and Y. E. Sagduyu, "DeepWiFi: Cognitive WiFi with deep learning," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 429–444, Feb. 2021.

[10] O. Omotere, J. Fuller, L. Qian, and Z. Han, "Spectrum occupancy prediction in coexisting wireless systems using deep learning," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, 2018, pp. 1–7.

[11] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for wireless resource management," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2017, pp. 1–6.

[12] T. S. Cousik, V. K. Shah, J. H. Reed, T. Erpek, and Y. E. Sagduyu, "Fast initial access with deep learning for beam prediction in 5G mmWave networks," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2021, pp. 664–669.

[13] N. Abuzainab, M. Alrabeiah, A. Alkhateeb, and Y. E. Sagduyu, "Deep learning for THz drones with flying intelligent surfaces: Beam and handoff prediction," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2021, pp. 1–6.

[14] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10.

[15] Y. Vorobeychik and M. Kantarcioglu, "Adversarial machine learning," *Synthesis Lectures Artif. Intell. Mach. Learn.*, vol. 12, no. 3, pp. 1–169, Dec. 2017.

[16] Y. Shi, Y. E. Sagduyu, K. Davaslioglu, and R. Levy, "Vulnerability detection and analysis in adversarial deep learning," in *Guide to Vulnerability Analysis for Computer Networks and Systems*. Cham, Switzerland: Springer, 2018, pp. 211–234.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, *arXiv:1412.6572*.

[18] N. H. Mahmood, H. Alves, O. A. López, M. Shehab, D. P. M. Osorio, and M. Latva-Aho, "Six key features of machine type communication in 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[19] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE Access*, vol. 6, pp. 32328–32338, 2018.

[20] G. Wikström *et al.*, "Challenges and technologies for 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[21] L. Zhang, Y.-C. Liang, and D. Niyato, "6G visions: Mobile ultra-broadband, super Internet-of-Things, and artificial intelligence," *China Commun.*, vol. 16, no. 8, pp. 1–14, Aug. 2019.

[22] X. Liao, L. Ding, and Y. Wang, "Secure machine learning, a brief overview," in *Proc. 5th Int. Conf. Secure Softw. Integr. Rel. Improvement Compan.*, 2011, pp. 26–29.

[23] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proc. ACM Symp. Inf. Comput. Commun. Security*, 2006, pp. 16–25.

[24] F. O. Olowononi, D. B. Rawat, and C. Liu, "Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 524–552, 1st Quart., 2021.

[25] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing connected autonomous vehicles: Challenges posed by adversarial machine learning and the way forward," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 998–1026, 2nd Quart., 2020.

[26] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[27] H. Xu *et al.*, "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, pp. 151–178, Mar. 2020.

[28] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3868–3880, Jun. 2022.

[29] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," in *Proc. 54th Annu. Conf. Inf. Sci. Syst. (CISS)*, 2020, pp. 1–6.

[30] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, "Adversarial attacks with multiple antennas against deep learning-based modulation classifiers," in *Proc. IEEE Globecom Workshops*, 2020, pp. 1–6.

[31] M. Z. Hameed, A. Gyorgy, and D. Gunduz, "The best defense is a good offense: Adversarial attacks to avoid modulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1074–1087, 2020.

[32] F. Restuccia *et al.*, "Generalized wireless adversarial deep learning," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiseML)*, 2020, pp. 49–54.

[33] M. DelVecchio, B. Flowers, and W. C. Headley, "Effects of forward error correction on communications aware evasion attacks," 2020, *arXiv:2005.13123*.

[34] Y. E. Sagduyu, Y. Shi, and T. Erpek, "Adversarial deep learning for over-the-air spectrum poisoning attacks," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 306–319, Feb. 2021.

[35] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of adversarial attacks in DNN-based modulation recognition," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2020, pp. 2469–2478.

[36] Y. E. Sagduyu *et al.*, "When wireless security meets machine learning: Motivation, challenges, and research directions," 2020, *arXiv:2001.08883*.

[37] D. Roy, T. Mukherjee, and M. Chatterjee, "Machine learning in adversarial RF environments," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 82–87, May 2019.

[38] D. Adesina, J. Bassey, and L. Qian, "Robust deep radio frequency spectrum learning for future wireless communications systems," *IEEE Access*, vol. 8, pp. 148528–148540, 2020.

[39] D. Adesina., J. Bassey, and L. Qian, "Practical radio frequency learning for future wireless communication systems," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2019, pp. 311–317.

[40] Y. Shi, K. Davaslioglu, Y. E. Sagduyu, W. C. Headley, M. Fowler, and G. Green, "Deep learning for RF signal classification in unknown and dynamic spectrum environments," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, 2019, pp. 1–10.

[41] J. Deogirikar and A. Vidhate, "Security attacks in IoT: A survey," in *Proc. Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, 2017, pp. 32–37.

[42] F. Shahzad, M. Pasha, and A. Ahmad, "A survey of active attacks on wireless sensor networks and their countermeasures," 2017, *arXiv:1702.07136*.

[43] S. Jaitly, H. Malhotra, and B. Bhushan, "Security vulnerabilities and countermeasures against jamming attacks in wireless sensor networks: A survey," in *Proc. Int. Conf. Comput. Commun. Electron. (Comptelix)*, 2017, pp. 559–564.

[44] S. Vadlamani, B. Eksioglu, H. Medal, and A. Nandi, "Jamming attacks on wireless networks: A taxonomic survey," *Int. J. Prod. Econ.*, vol. 172, pp. 76–94, Feb. 2016.

[45] Y. Wu, A. Khisti, C. Xiao, G. Caire, K.-K. Wong, and X. Gao, "A survey of physical layer security techniques for 5G wireless networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 679–695, Apr. 2018.

[46] I. Stellios, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, "A survey of IoT-enabled cyberattacks: Assessing attack paths to critical infrastructures and services," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3453–3495, 4th Quart., 2018.

[47] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A survey on wireless security: Technical challenges, recent advances, and future trends," *Proc. IEEE*, vol. 104, no. 9, pp. 1727–1765, Sep. 2016.

[48] M. Arif, G. Wang, M. Z. A. Bhuiyan, T. Wang, and J. Chen, "A survey on security attacks in VANETs: Communication, applications and challenges," *Veh. Commun.*, vol. 19, Oct. 2019, Art. no. 100179.

[49] P. Dewal, G. S. Narula, V. Jain, and A. Baliyan, "Security attacks in wireless sensor networks: A survey," in *Cyber Security*, M. U. Bokhari, N. Agrawal, and D. Saini, Eds. Singapore: Springer, 2018, pp. 47–58.

[50] M. Parashar, A. Poonia, and K. Satish, "A survey of attacks and their Mitigations in software defined networks," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, 2019, pp. 1–8.

[51] Y. E. Sagduyu, Y. Shi, and T. Erpek, "IoT network security from the perspective of adversarial deep learning," in *Proc. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, 2019, pp. 1–9.

[52] Y. Shi, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, Z. Lu, and J. H. Li, "Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2018, pp. 1–6.

[53] Y. Shi, T. Erpek, Y. E. Sagduyu, and J. H. Li, "Spectrum data poisoning with adversarial deep learning," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2018, pp. 407–412.

[54] M. Sadeghi and E. G. Larsson, "Physical adversarial attacks against end-to-end Autoencoder communication systems," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 847–850, May 2019.

[55] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1102–1113, 2020.

[56] F. Restuccia *et al.*, "Hacking the waveform: Generalized wireless adversarial deep learning," 2020, *arXiv:2005.02270*.

[57] S. Kokalj-Filipovic, R. Miller, and J. Morman, "Targeted adversarial examples against RF deep classifiers," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiseML)*, 2019, pp. 6–11.

[58] M. DelVecchio, V. Arndorfer, and W. C. Headley, "Investigating a spectral deception loss metric for training machine learning-based evasion attacks," 2020, *arXiv:2005.13124*.

[59] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, "Channel effects on surrogate models of adversarial attacks against wireless signal classifiers," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[60] C. Zhong, F. Wang, M. C. Gursoy, and S. Velipasalar, "Adversarial jamming attacks on deep reinforcement learning based dynamic multichannel access," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2020, pp. 1–6.

[61] F. Wang, C. Zhong, M. C. Gursoy, and S. Velipasalar, "Adversarial jamming attacks and defense strategies via adaptive deep reinforcement learning," 2020, *arXiv:2007.06055*.

[62] Y. Shi, Y. E. Sagduyu, T. Erpek, and M. C. Gursoy, "How to attack and defend 5G radio access network slicing with reinforcement learning," 2021, *arXiv:2101.05768*.

[63] Y. Shi and Y. E. Sagduyu, "Adversarial machine learning for flooding attacks on 5G radio access network slicing," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2021, pp. 1–6.

[64] S. Bhambri, S. Muku, A. Tulasi, and A. B. Buduru, "A survey of black-box adversarial attacks on computer vision models," 2020, *arXiv:1912.01667*.

[65] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.

[66] I. Tomić and J. A. McCann, "A survey of potential security issues in existing wireless sensor network protocols," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1910–1923, Dec. 2017.

[67] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, Dec. 2018.

[68] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Security Artif. Intell.*, 2011, pp. 43–58.

[69] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," 2018, *arXiv:1810.00069*.

[70] E. Nowroozi, A. Dehghantanha, R. M. Parizi, and K.-K. R. Choo, "A survey of machine learning techniques in adversarial image forensics," *Comput. Security*, vol. 100, Jan. 2021, Art. no. 102092.

[71] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey towards the defender's perspective," 2020, *arXiv:2009.03728*.

[72] A. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM Comput. Surveys*, vol. 53, no. 3, p. 66, Jun. 2020.

[73] L. Sun *et al.*, "Adversarial attack and defense on graph data: A survey," 2020, *arXiv:1812.10528*.

[74] Y. Zou, J. Zhu, L. Yang, Y. Liang, and Y. Yao, "Securing physical-layer communications for cognitive radio networks," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 48–54, Sep. 2015.

[75] Y. E. Sagduyu, "Securing cognitive radio networks with dynamic trust against spectrum sensing data falsification," in *Proc. IEEE Mil. Commun. Conf.*, 2014, pp. 235–241.

[76] Y. E. Sagduyu, R. A. Berry, and A. Ephremides, "Jamming games in wireless networks with incomplete information," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 112–118, Aug. 2011.

[77] Y. E. Sagduyu, R. A. Berry, and A. Ephremides, "Wireless jamming attacks under dynamic traffic uncertainty," in *Proc. Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, 2010, pp. 303–312.

[78] S. Rajendran, W. Meert, V. Lenders, and S. Pollin, "SAIFE: Unsupervised wireless spectrum anomaly detection with interpretable features," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, 2018, pp. 1–9.

[79] N. Van Huynh, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, "'Jam me if you can': Defeating jammer with deep Dueling neural network architecture and ambient backscattering augmented communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2603–2620, Nov. 2019.

[80] N. Abuzainab *et al.*, "QoS and jamming-aware wireless networking using deep reinforcement learning," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2019, pp. 610–615.

[81] Y. Zhou, M. Kantarcioglu, and B. Xi, "A survey of game theoretic approach for adversarial machine learning," *Wiley Interdiscipl. Rev. Data Min. Knowl. Disc.*, vol. 9, no. 3, 2019, Art. no. e1259.

[82] X. Wang, J. Li, X. Kuang, Y. Tan, and J. Li, "The security of machine learning in an adversarial setting: A survey," *J. Parallel Distrib. Comput.*, vol. 130, pp. 12–23, Aug. 2019.

[83] G. Li, P. Zhu, J. Li, Z. Yang, N. Cao, and Z. Chen, "Security matters: A survey on adversarial machine learning," 2018, *arXiv:1810.07339*.

[84] M. Ozdag, "Adversarial attacks and defenses against deep neural networks: A survey," *Procedia Comput. Sci.*, vol. 140, pp. 152–161, Nov. 2018.

[85] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE Access*, vol. 6, pp. 12103–12117, 2018.

[86] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A taxonomy and survey of attacks against machine learning," *Comput. Sci. Rev.*, vol. 34, Nov. 2019, Art. no. 100199.

[87] S. H. Silva and P. Najafirad, "Opportunities and challenges in deep learning adversarial robustness: A survey," 2020, *arXiv:2007.00753*.

[88] R. R. Wiyatno, A. Xu, O. Dia, and A. de Berker, "Adversarial examples in modern machine learning: A review," 2019, *arXiv:1911.05268*.

[89] Q. Liu, J. Guo, C. Wen, and S. Jin, "Adversarial attack on DL-based massive MIMO CSI feedback," *J. Commun. Netw.*, vol. 22, no. 3, pp. 230–235, Jun. 2020.

[90] M. Kulin, T. Kazaz, E. De Poorter, and I. Moerman, "A survey on machine learning-based performance improvement of wireless networks: PHY, MAC and network layer," *Electronics*, vol. 10, no. 3, p. 318, 2021.

[91] S. Zheng *et al.*, "Big data processing architecture for radio signals empowered by deep learning: Concept, experiment, applications and challenges," *IEEE Access*, vol. 6, pp. 55907–55922, 2018.

[92] L. Qian, J. Zhu, and S. Zhang, "Survey of wireless big data," *J. Commun. Inf. Netw.*, vol. 2, no. 1, pp. 1–18, 2017. [Online]. Available: http://dx.doi.org/10.1007/s41650-017-0001-2

[93] L. Dai, R. Jiao, F. Adachi, H. V. Poor, and L. Hanzo, "Deep learning for wireless communications: An emerging interdisciplinary paradigm," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 133–139, Aug. 2020.

[94] J. G. Proakis, *Digital Communications 5th Edition*. Boston, MA, USA: McGraw Hill, 2007.

[95] M. Kulin, T. Kazaz, I. Moerman, and E. D. Poorter, "End-to-end learning from spectrum data: A deep learning approach for wireless signal identification in spectrum monitoring applications," *IEEE Access*, vol. 6, pp. 18484–18501, 2018.

[96] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[97] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, 4th Quart., 2019.

[98] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, "A survey of machine learning techniques applied to self-Organizing cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2392–2431, 4th Quart., 2017.

[99] D. C. Nguyen *et al.*, "Enabling AI in future wireless networks: A data life cycle perspective," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 553–595, 1st Quart., 2021.

[100] J. Jagannath, N. Polosky, A. Jagannath, F. Restuccia, and T. Melodia, "Machine learning for wireless communications in the Internet of Things: A comprehensive survey," *Ad Hoc Netw.*, vol. 93, Oct. 2019, Art. no. 101913.

[101] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.

[102] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[103] G. Y. Li *et al.*, "Series editorial: The second issue of the series on machine learning in communications and networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1855–1857, Jul. 2021.

[104] T. Huynh-The, C.-H. Hua, V.-S. Doan, Q.-V. Pham, T.-V. Nguyen, and D.-S. Kim, "Deep learning for constellation-based modulation classification under Multipath fading channels," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, 2020, pp. 300–304.

[105] Y. Kumar, M. Sheoran, G. Jajoo, and S. K. Yadav, "Automatic modulation classification based on constellation density using deep learning," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1275–1278, Jun. 2020.

[106] S. Peng *et al.*, "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 718–727, Mar. 2019.

[107] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," 2016, *arXiv:1602.04105*.

[108] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.

[109] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 433–445, Sep. 2018.

[110] S. Peng, S. Sun, and Y.-D. Yao, "A survey of modulation classification using deep learning: Signal representation and data preprocessing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 14, 2021, doi: 10.1109/TNNLS.2021.3085433.

[111] R. Zhou, F. Liu, and C. W. Gravelle, "Deep learning for modulation recognition: A survey with a demonstration," *IEEE Access*, vol. 8, pp. 67366–67376, 2020.

[112] S. Zheng, S. Chen, P. Qi, H. Zhou, and X. Yang, "Spectrum sensing based on deep learning classification for cognitive radios," *China Commun.*, vol. 17, no. 2, pp. 138–148, Feb. 2020.

[113] J. Xie, J. Fang, C. Liu, and X. Li, "Deep learning-based spectrum sensing in cognitive radio: A CNN-LSTM approach," *IEEE Commun. Lett.*, vol. 24, no. 10, pp. 2196–2200, Oct. 2020.

[114] W. Lee, M. Kim, and D.-H. Cho, "Deep cooperative sensing: Cooperative spectrum sensing based on convolutional neural networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3005–3009, Mar. 2019.

[115] W. M. Lees, A. Wunderlich, P. J. Jeavons, P. D. Hale, and M. R. Souryal, "Deep learning classification of 3.5-GHz band spectrograms with applications to spectrum sensing," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 2, pp. 224–236, Jun. 2019.

[116] C. Liu, J. Wang, X. Liu, and Y.-C. Liang, "Deep CM-CNN for spectrum sensing in cognitive radio," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2306–2321, Oct. 2019.

[117] J. Gao, X. Yi, C. Zhong, X. Chen, and Z. Zhang, "Deep learning for spectrum sensing," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1727–1730, Dec. 2019.

[118] J. Xie, C. Liu, Y.-C. Liang, and J. Fang, "Activity pattern aware spectrum sensing: A CNN-based deep learning approach," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 1025–1028, Jun. 2019.

[119] Y. Arjoune and N. Kaabouch, "A comprehensive survey on spectrum sensing in cognitive radio networks: Recent advances, new challenges, and future research directions," *Sensors*, vol. 19, no. 1, p. 126, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/1/126

[120] V. Ramani and S. K. Sharma, "Cognitive radios: A survey on spectrum sensing, security and spectrum handoff," *China Commun.*, vol. 14, no. 11, pp. 185–208, Nov. 2017.

[121] M. Eisen and A. Ribeiro, "Optimal wireless resource allocation with random edge graph neural networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 2977–2991, Apr. 2020.

[122] Y. Zhou, Z. M. Fadlullah, B. Mao, and N. Kato, "A deep-learning-based radio resource assignment technique for 5G ultra dense networks," *IEEE Netw.*, vol. 32, no. 6, pp. 28–34, Nov./Dec. 2018.

[123] J. Guo and C. Yang, "Predictive resource allocation with deep learning," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, 2018, pp. 1–7.

[124] J. Gao, M. R. A. Khandaker, F. Tariq, K.-K. Wong, and R. T. Khan, "Deep neural network based resource allocation for V2X communications," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, 2019, pp. 1–5.

[125] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2017, pp. 1–6.

[126] K. I. Ahmed, H. Tabassum, and E. Hossain, "Deep learning for radio resource allocation in multi-cell networks," *IEEE Netw.*, vol. 33, no. 6, pp. 188–195, Nov./Dec. 2019.

[127] H. Zhang, M. Feng, K. Long, G. K. Karagiannidis, and A. Nallanathan, "Artificial intelligence-based resource allocation in ultradense networks: Applying event-triggered Q-learning algorithms," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 56–63, Dec. 2019.

[128] U. Challita, L. Dong, and W. Saad, "Deep learning for proactive resource allocation in LTE-U networks," in *Proc. 23th Eur. Wireless Conf.*, 2017, pp. 1–6.

[129] I. Ahmad *et al.*, "Machine learning meets communication networks: Current trends and future challenges," *IEEE Access*, vol. 8, pp. 223418–223460, 2020.

[130] L. Zhang, Y.-C. Liang, and D. Niyato, "6G visions: Mobile ultra-broadband, super Internet-of-Things, and artificial intelligence," *China Commun.*, vol. 16, no. 8, pp. 1–14, Aug. 2019.

[131] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 3rd Quart., 2020.

[132] 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, Feb. 2019.

[133] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020.

[134] S. S. Islam, S. Rahman, M. M. Rahman, E. K. Dey, and M. Shoyaib, "Application of deep learning to computer vision: A comprehensive study," in *Proc. 5th Int. Conf. Inform., Electron. Vis. (ICIEV)*, 2016, pp. 592–597.

[135] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349.

[136] M. Hassaballah and A. I. Awad, *Deep Learning in Computer Vision: Principles and Applications*. Boca Raton, FL, USA: CRC Press, 2020.

[137] J. Ni, T. Young, V. Pandelea, F. Xue, V. Adiga, and E. Cambria, "Recent advances in deep learning based dialogue systems: A systematic survey," 2021, *arXiv:2105.04387*.

[138] G. Weisz, P. Budzianowski, P.-H. Su, and M. Gašić, "Sample efficient deep reinforcement learning for dialogue systems with large action spaces," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2083–2097, Nov. 2018.

[139] M. Korpusik and J. Glass, "Deep learning for database mapping and asking clarification questions in dialogue systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1321–1334, Aug. 2019.

[140] Y.-N. Chen, A. Celikyilmaz, and D. Hakkani-Tur, "Deep learning for dialogue systems," in *Proc. 27th Int. Conf. Comput. Linguist. Tutorial Abstracts*, 2018, pp. 25–31.

[141] P. Beckmann, M. Kegler, H. Saltini, and M. Cernak, "Speech-VGG: A deep feature extractor for speech processing," 2019, *arXiv:1910.09909*.

[142] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," 2018, *arXiv:1804.00015*.

[143] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, Aug. 2017.

[144] R. Haeb-Umbach *et al.*, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124, Nov. 2019.

[145] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2018, pp. 2847–2856. [Online]. Available: https://doi.org/10.1145/3219819.3220078

[146] L. Chen *et al.*, "A survey of adversarial learning on graphs," 2020, *arXiv:2003.05730*.

[147] X. Zhang and M. Zitnik, "GNNGuard: Defending graph neural networks against adversarial attacks," in *Proc. Neural Inf. Process. Syst.*, 2020, pp. 1–13.

[148] J. Xu, J. Chen, S. You, Z. Xiao, Y. Yang, and J. Lu, "Robustness of deep learning models on graphs: A survey," *AI Open*, vol. 2, pp. 69–78, Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000139

[149] B. Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in Bioinformatics and computational biology," *Front. Genet.*, vol. 10, p. 214, Mar. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2019.00214

[150] A. Yazdani, L. Lu, M. Raissi, and G. E. Karniadakis, "Systems biology informed deep learning for inferring parameters and hidden dynamics," *PLoS Comput. Biol.*, vol. 16, no. 11, 2020, Art. no. e1007575.

[151] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: New computational modelling techniques for genomics," *Nat. Rev. Genet.*, vol. 20, no. 7, pp. 389–403, 2019.

[152] L. Rampasek and A. Goldenberg, "TensorFlow: Biology's gateway to deep learning?" *Cell Syst.*, vol. 2, no. 1, pp. 12–14, 2016.

[153] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[154] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–10, 2018.

[155] B. Norgeot *et al.*, "Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis," *JAMA Netw. Open*, vol. 2, no. 3, 2019, Art. no. e190606.

[156] D. Liu, T. Miller, R. Sayeed, and K. D. Mandl, "FADL: Federated-autonomous deep learning for distributed electronic health record," 2018, *arXiv:1811.11400*.

[157] A. B. Levine, C. Schlosser, J. Grewal, R. Coope, S. J. Jones, and S. Yip, "Rise of the machines: Advances in deep learning for cancer diagnosis," *Trends Cancer*, vol. 5, no. 3, pp. 157–169, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405803319300184

[158] A. G. C. Pacheco and R. A. Krohling, "Recent advances in deep learning applied to skin cancer detection," 2019, *arXiv:1912.03280*.

[159] Y. Xiao, J. Wu, and Z. Lin, "Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104540.

[160] U. Kose and J. Alzubi, *Deep Learning for Cancer Diagnosis*. Singapore: Springer, 2021.

[161] K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi, and T. Sato, "Deep-learning-empowered breast cancer auxiliary diagnosis for 5GB remote e-health," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 54–61, Jun. 2021.

[162] W. Sun, B. Zheng, and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," in *Proc. Med. Imag. Comput.-Aided Diagnosis*, 2016, Art. no. 97850Z.

[163] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep learning advances in computer vision with 3D data: A survey," *ACM Comput. Surveys*, vol. 50, no. 2, p. 20, Apr. 2017. [Online]. Available: https://doi.org/10.1145/3042064

[164] C. R. Qi, "Deep learning on 3D data," in *3D Imaging, Analysis and Applications*. Cham, Switzerland: Springer, 2020, pp. 513–566.

[165] X. Chai, G. Tang, S. Wang, K. Lin, and R. Peng, "Deep learning for irregularly and regularly missing 3-D data reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6244–6265, Jul. 2021.

[166] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.

[167] X. Zhou and S. Prasad, "Advances in deep learning for hyperspectral image analysis–addressing challenges arising in practical imaging scenarios," in *Proc. Adv. Comput. Vis. Pattern Recognit.*, 2020, pp. 117–140.

[168] A. Signoroni, M. Savardi, A. Baronio, and S. Benini, "Deep learning meets hyperspectral image analysis: A multidisciplinary review," *J. Imag.*, vol. 5, no. 5, p. 52, 2019.

[169] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[170] X. Yang, Y. Ye, X. Li, R. Y. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[171] Z. Gong, P. Zhong, and W. Hu, "Statistical loss and analysis for deep learning in hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 322–333, Jan. 2021.

[172] S. Kokalj-Filipovic, R. Miller, N. Chang, and C. L. Lau, "Mitigation of adversarial examples in RF deep classifiers utilizing AutoEncoder pre-training," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, 2019, pp. 1–6.

[173] K. Davaslioglu and Y. E. Sagduyu, "Trojan attacks on wireless signal classification with adversarial machine learning," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, 2019, pp. 1–6.

[174] S. Bair, M. DelVecchio, B. Flowers, A. J. Michaels, and W. C. Headley, "On the limitations of targeted adversarial evasion attacks against deep learning enabled modulation recognition," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiseML)*, 2019, pp. 25–30.

[175] D. Ke, Z. Huang, X. Wang, and L. Sun, "Application of adversarial examples in communication modulation classification," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, 2019, pp. 877–882.

[176] M. Z. Hameed, A. György, and D. Gündüz, "Communication without interception: Defense against modulation detection," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, 2019, pp. 1–5.

[177] S. Kokalj-Filipovic and R. Miller, "Adversarial examples in RF learning: Detection of the attack and its physical robustness," 2019, *arXiv:1902.06044*.

[178] S. Karunaratne, E. Krijestorac, and D. Cabric, "Penetrating RF fingerprinting-based authentication with a generative adversarial attack," 2020, *arXiv:2011.01538*.

[179] F. Wang, C. Zhong, M. C. Gursoy, and S. Velipasalar, "Defense strategies against adversarial jamming attacks via deep reinforcement learning," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, 2020, pp. 1–6.

[180] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Over-the-air membership inference attacks as privacy threats for deep learning-based wireless signal classifiers," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiseML)*, 2020, pp. 61–66.

[181] Y. Shi and Y. Sagduyu, "Membership inference attack and defense for wireless signal classifiers with deep learning," *IEEE Trans. Mobile Comput.*, early access, Feb. 7, 2022, doi: 10.1109/TMC.2022.3148690.

[182] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "How to make 5G communications 'invisible': Adversarial machine learning for wireless privacy," in *Proc. 54th Asilomar Conf. Signals Syst. Comput.*, 2020, pp. 763–767.

[183] Y.-J. Cao *et al.*, "Recent advances of generative adversarial networks in computer vision," *IEEE Access*, vol. 7, pp. 14985–15006, 2019.

[184] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155161–155196, 2021.

[185] Z. Zou, S. Lei, T. Shi, Z. Shi, and J. Ye, "Deep adversarial decomposition: A unified framework for separating superimposed images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12806–12816.

[186] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5814–5824.

[187] H. Wang, D. Gong, Z. Li, and W. Liu, "Decorrelated adversarial learning for age-invariant face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3527–3536.

[188] Y. Zhong and W. Deng, "Adversarial learning with margin-based triplet embedding regularization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6549–6558.

[189] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning models," 2018, *arXiv:1707.08945*.

[190] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[191] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," 2021, *arXiv:2102.01356*.

[192] A. Kurakin, D. Boneh, F. Tramèr, I. Goodfellow, N. Papernot, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. ICLR*, 2018, pp. 1–20. [Online]. Available: https://openreview.net/pdf?id=rkZvSe-RZ

[193] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 4312–4321. [Online]. Available: https://doi.org/10.24963/ijcai.2021/591

[194] M. Qiu and H. Qiu, "Review on image processing based adversarial example defenses in computer vision," in *Proc. IEEE 6th Int. Conf. Big Data Security Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. (HPSC) IEEE Int. Conf. Intell. Data Security (IDS)*, 2020, pp. 94–99.

[195] F. Feng, X. He, J. Tang, and T.-S. Chua, "Graph adversarial training: Dynamically regularizing based on graph structure," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2493–2504, Jun. 2021.

[196] Y. Qiao, X. Luo, C. Li, H. Tian, and J. Ma, "Heterogeneous graph-based joint representation learning for users and POIs in location-based social network," *Inf. Process. Manage.*, vol. 57, Mar. 2020, Art. no. 102151.

[197] Q. Tan, N. Liu, and X. Hu, "Deep representation learning for social network analysis," *Front. Big Data*, vol. 2, p. 2, Apr. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdata.2019.00002

[198] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," 2018, *arXiv:1706.02216*.

[199] A. Abusnaina, A. Khormali, H. Alasmary, J. Park, A. Anwar, and A. Mohaisen, "Adversarial learning attacks on graph-based IoT Malware detection systems," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, 2019, pp. 1296–1305.

[200] H. Dai *et al.*, "Adversarial attack on graph structured data," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1115–1124.

[201] Z. Zhang, Z. Zhang, Y. Zhou, Y. Shen, R. Jin, and D. Dou, "Adversarial attacks on deep graph matching," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20834–20851.

[202] S. Pan, R. Hu, S.-F. Fung, G. Long, J. Jiang, and C. Zhang, "Learning graph embedding with adversarial training methods," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2475–2487, Jun. 2020.

[203] H. Wang *et al.*, "GraphGAN: Graph representation learning with generative adversarial nets," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 2508–2515.

[204] Y. Zhang, S. Khan, and M. Coates, "Comparing and detecting adversarial attacks for graph deep learning," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–7.

[205] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Comput. Surveys*, vol. 54, no. 3, p. 62, Apr. 2022. [Online]. Available: https://doi.org/10.1145/3439726

[206] J. Kim, S. Jang, E. Park, and S. Choi, "Text classification using capsules," *Neurocomputing*, vol. 376, pp. 214–221, Feb. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219314092

[207] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," 2018, *arXiv:1801.01944*.

[208] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.

[209] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," 2021, *arXiv:1605.07725*.

[210] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "HotFlip: White-box adversarial examples for text classification," 2018, *arXiv:1712.06751*.

[211] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 4208–4215.

[212] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," 2018, *arXiv:1804.06059*.

[213] S. Samanta and S. Mehta, "Towards crafting text adversarial samples," 2017, *arXiv:1707.02812*.

[214] Q. Lei, L. Wu, P.-Y. Chen, A. Dimakis, I. S. Dhillon, and M. J. Witbrock, "Discrete adversarial attacks and submodular optimization with applications to text classification," in *Proc. Mach. Learn. Syst.*, 2019, pp. 146–165.

[215] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, p. 24, 2020. [Online]. Available: https://doi.org/10.1145/3374217

[216] Y. Zhou, X. Zheng, C.-J. Hsieh, K.-W. Chang, and X. Huang, "Defense against adversarial attacks in NLP via dirichlet neighborhood ensemble," 2020, *arXiv:2006.11627*.

[217] V. Araujo, A. Carvallo, C. Aspillaga, and D. Parra, "On adversarial examples for biomedical NLP tasks," 2020, *arXiv:2004.11157*.

[218] D. Miao *et al.*, "Simple contrastive representation adversarial learning for NLP tasks," 2021, *arXiv:2111.13301*.

[219] X. Han, T. Baldwin, and T. Cohn, "Decoupling adversarial training for fair NLP," in *Proc. FINDINGS*, 2021, pp. 471–477.

[220] B. Alshemali and J. Kalita, "Improving the reliability of deep neural networks in NLP: A review," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105210, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119305428

[221] I. Alsmadi *et al.*, "Adversarial attacks and defenses for social network text processing applications: Techniques, challenges and future research directions," 2021, *arXiv:2110.13980*.

[222] W. Wang, R. Wang, L. Wang, Z. Wang, and A. Ye, "Towards a robust deep neural network against adversarial texts: A survey," *IEEE Trans. Knowl. Data Eng.*, early access, Oct. 4, 2021, doi: 10.1109/TKDE.2021.3117608.

[223] R. Sahay, C. G. Brinton, and D. J. Love, "Frequency-based automated modulation classification in the presence of adversaries," 2020, *arXiv:2011.01132*.

[224] A. Berian, K. Staab, N. Teku, G. Ditzler, T. Bose, and R. Tandon, "Adversarial filters for secure modulation classification," 2020, *arXiv:2008.06785*.

[225] B. Kim, T. Erpek, Y. E. Sagduyu, and S. Ulukus, "Covert communications via adversarial machine learning and reconfigurable intelligent surfaces," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 411–416.

[226] Y. E. Sagduyu, R. A. Berry, and D. Guo, "Throughput and stability for relay-assisted wireless broadcast with network coding," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1506–1516, Aug. 2013.

[227] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 1, pp. 2–14, Mar. 2019.

[228] Z. Luo, S. Zhao, Z. Lu, J. Xu, and Y. E. Sagduyu, "When attackers meet AI: Learning-empowered attacks in cooperative spectrum sensing," *IEEE Trans. Mobile Comput.*, vol. 21, no. 5, pp. 1892–1908, May 2022.

[229] Z. Luo, S. Zhao, Z. Lu, Y. E. Sagduyu, and J. Xu, "Adversarial machine learning based partial-model attack in IoT," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiseML)*, 2020, pp. 13–18.

[230] S. Kokalj-Filipovic, R. Miller, and G. Vanhoy, "Adversarial examples in RF deep learning: Detection and physical robustness," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, 2019, pp. 1–5.

[231] H. Kwon, Y. Kim, H. Yoon, and D. Choi, "Fooling a neural network in military environments: Random untargeted adversarial example," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2018, pp. 456–461.

[232] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and privacy in machine learning," in *Proc. IEEE Eur. Symp. Security Privacy (EuroS P)*, 2018, pp. 399–414.

[233] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. USENIX Security Symp.*, 2016, pp. 1–18.

[234] Y. Shi, Y. E. Sagduyu, and A. Grushin, "How to steal a machine learning classifier with deep learning," in *Proc. IEEE Int. Symp. Technol. Homeland Security (HST)*, 2017, pp. 1–5.

[235] Y. Shi, Y. E. Sagduyu, K. Davaslioglu, and J. H. Li, "Active deep learning attacks under strict rate limitations for online API calls," in *Proc. IEEE Int. Symp. Technol. Homeland Security (HST)*, 2018, pp. 1–6.

[236] L. Pi, Z. Lu, Y. Sagduyu, and S. Chen, "Defending active learning against adversarial inputs in automated document classification," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, 2016, pp. 257–261.

[237] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[238] Y. Shi, Y. E. Sagduyu, K. Davaslioglu, and J. H. Li, "Generative adversarial networks for black-box API attacks with limited training data," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, 2018, pp. 453–458.

[239] K. Davaslioglu and Y. E. Sagduyu, "Generative adversarial learning for spectrum sensing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–6.

[240] E. Ayanoglu, K. Davaslioglu, and Y. E. Sagduyu, "Machine learning in NextG networks via generative adversarial networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 480–501, Jun. 2022.

[241] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Security (AsiaCCS)*, 2017, pp. 506–519.

[242] B. Manoj, M. Sadeghi, and E. G. Larsson, "Adversarial attacks on deep learning based power allocation in a massive MIMO network," 2021, *arXiv:2101.12090*.

[243] T. Hou, T. Wang, Z. Lu, Y. Liu, and Y. Sagduyu, "Undermining deep learning based channel estimation via adversarial wireless signal fabrication," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiSec-WiseML)*, May 2022, pp. 63–68.

[244] T. Hou *et al.*, "MUSTER: Subverting user selection in MU-MIMO networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2022, pp. 140–149.

[245] B. Kim, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond," in *Proc. IEEE Stat. Signal Process. Workshop*, 2021, pp. 590–594.

[246] W. Xu, W. Trappe, Y. Zhang, and T. Wood, "The feasibility of launching and detecting jamming attacks in wireless networks," in *Proc. ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2005, pp. 46–57.

[247] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 1467–1474.

[248] Y. Shi and Y. E. Sagduyu, "Evasion and causative attacks with adversarial deep learning," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, 2017, pp. 243–248.

[249] Z. Luo, S. Zhao, R. Duan, Z. Lu, Y. E. Sagduyu, and J. Xu, "Low-cost influence-limiting defense against adversarial machine learning attacks in cooperative spectrum sensing," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiseML)*, 2021, pp. 1–6.

[250] Y. He, G. Meng, K. Chen, X. Hu, and J. He, "Towards security threats of deep learning systems: A survey," *IEEE Trans. Softw. Eng.*, vol. 48, no. 5, pp. 1743–1770, May 2022.

[251] B. Wang *et al.*, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Security Privacy (SP)*, 2019, pp. 707–723.

[252] B. Chen *et al.*, "Detecting backdoor attacks on deep neural networks by activation clustering," 2018, *arXiv:1811.03728*.

[253] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*.

[254] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1765–1773.

[255] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9185–9193.

[256] Y. Shi, Y. E. Sagduyu, and T. Erpek, "Reinforcement learning for dynamic resource optimization in 5G radio access network slicing," in *Proc. IEEE Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, 2020, pp. 1–6.

[257] Y. Shi, P. Rahimzadeh, M. Costa, T. Erpek, and Y. E. Sagduyu, "Deep reinforcement learning for 5G radio access network slicing with spectrum coexistence," TechRxiv, 2021.

[258] O. A. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "Survey of automatic modulation classification techniques: Classical approaches and new trends," *IET Commun.*, vol. 1, no. 2, pp. 137–156, Apr. 2007.

[259] T. J. O'Shea and N. West, "Radio machine learning dataset generation with GNU radio," in *Proc. GNU Radio Conf.*, 2016, pp. 1–6.

[260] M. Z. Hameed. "New quality measures for adversarial attacks with applications to secure communication." Jul. 2020. [Online]. Available: https://spiral.imperial.ac.uk:8443/handle/10044/1/82214

[261] J. Yi and A. E. Gamal, "Gradient-based adversarial deep modulation classification with data-driven subsampling," 2021, *arXiv:2104.06375*.

[262] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against DNN-based wireless communication systems," 2021, *arXiv:2102.00918*.

[263] A. Saeed, K. A. Harras, E. Zegura, and M. Ammar, "Local and low-cost white space detection," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, 2017, pp. 503–516.

[264] Y. E. Sagduyu, T. Erpek, and Y. Shi, "Adversarial machine learning for 5G communications security," in *Proc. IEEE Game Theory Mach. Learn. Cyber Security*, 2021, pp. 270–288.

[265] F. Wang, M. C. Gursoy, and S. Velipasalar, "Adversarial reinforcement learning in dynamic channel access and power control," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2021, pp. 1–6.

[266] L. Sanguinetti, A. Zappone, and M. Debbah, "Deep learning power allocation in massive MIMO," 2019, *arXiv:1812.03640*.

[267] B. Kim, Y. Shi, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial attacks against deep learning based power control in wireless communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM) Workshops*, 2021, pp. 1–6.

[268] T. Hou, T. Wang, Z. Lu, Y. Liu, and Y. Sagduyu, "IoTGAN: GAN powered camouflage against machine learning based IoT device identification," in *Proc. IEEE Int. Symp. Dynamic Spectr. Access Netw. (DySPAN)*, 2021, pp. 280–287.

[269] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network for wireless signal spoofing," in *Proc. ACM Workshop Wireless Security Mach. Learn. (WiseML)*, 2019, pp. 55–60.

[270] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network in the air: Deep adversarial learning for wireless signal spoofing," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 294–303, Mar. 2021.

[271] Y. Shi and Y. E. Sagduyu, "Jamming attacks on federated learning in wireless networks," 2022, *arxiv:2201.05172*.

[272] Y. Shi, Y. E. Sagduyu, and T. Erpek, "Federated learning for distributed spectrum sensing in NextG communication networks," in *Proc. Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl. IV*, 2022, pp. 472–478.

[273] L. Liu *et al.*, "The COST 2100 MIMO channel model," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, Dec. 2012.

[274] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," 2019, *arXiv:1902.02918*.

[275] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–15.

[276] T. Hou, Z. Qu, T. Wang, Z. Lu, and Y. Liu, "ProTO: Proactive topology obfuscation against adversarial network topology inference," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 1598–1607.

[277] "Genesys lab ML datasets." Genesys Lab. 2020. [Online]. Available: https://genesys-lab.org/mldatasets

[278] A. Al-shawabka, F. Restuccia, S. D'Oro, and T. Melodia, "Massive-scale I/Q datasets for WiFi radio fingerprinting," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107566.

[279] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *Proc. Inf. Theory Appl. Workshop (ITA)*, 2019, pp. 1–8.

[280] M. Ezuma, F. Erden, C. K. Anjinappa, O. Ozdemir, and I. Guvenc. "Drone remote controller RF signal dataset." 2020. [Online]. Available: https://dx.doi.org/10.21227/ss99-8d56

[281] J. Breen *et al.*, "POWDER: Platform for open wireless data-driven experimental research," in *Proc. 14th Int. Workshop Wireless Netw. Testbeds, Exp. Eval. Characterization (WiNTECH)*, Sep. 2020, pp. 17–24.

[282] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," 2019, *arXiv:1905.02175*.

[283] R. Rai and C. K. Sahu, "Driven by data or derived through physics? A review of hybrid physics guided machine learning techniques with cyber-physical system (CPS) focus," *IEEE Access*, vol. 8, pp. 71050–71073, 2020.