GAN-Based Detection of Adversarial EM Signal Waveforms

Athanasios Gkelias
Electrical and Electronic Engineering Dept.
Imperial College London
London, UK
a.gkelias@imperial.ac.uk

Kin K. Leung
Electrical and Electronic Engineering Dept.
Imperial College London
London, UK
kin.leung@imperial.ac.uk

MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM) | 978-1-6654-8534-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/MILCOM55135.2022.10017943

***Abstract*—Detection of unauthorised or malicious electromagnetic (EM) transmissions in the wireless spectrum is highly important in both military and commercial systems. In military wireless networks, and particularly in congested EM environments, the detection of unknown radar or communication waveforms can lead to timely identification of potentially adversarial transmissions or intruders in the area. On the other hand, in cognitive radio networks the identification of unauthorised communication waveforms can prevent and mitigate security threats, such as Primary User Emulation (PUE) attacks. However, data of such waveforms are usually of insignificant size to be effectively modelled or even there are no prior data available since they appear for the first time, which makes their timely detection particularly difficult.**

In this paper, we present a Generative Adversarial Network (GAN) based system which trains on available (presumably friendly) EM signals to detect any previously unseen types of EM waveforms, which can be potentially characterised as unauthorised or malicious. The proposed system is successfully trained and tested on a synthetic dataset comprising different pulsed radar and communication modulated signals impaired with Rician multipath fading, AWGN and random clock offset, resulting in center frequency offset and sampling time drift, and it was shown to successfully detect any previously unseen types of EM waveforms even in low SNR.

***Index Terms*—Anomaly Detection, Generative Adversarial Networks, Wireless Communications, Radar Waveforms, Wireless Security.**

I. Introduction

The electromagnetic (EM) environment is becoming increasingly congested and a major challenge in both military and civilian wireless networks is the detection, classification and management of radiowave interference. Particularly in military systems and tactical networks, the detection of unknown (i.e., previously unseen) radar waveforms, radio communication signals or other EM transmissions is highly important since it may indicate the presence of hostile communication, radar or sensing devices in the close area. On the other hand, in cognitive radio networks, one of the most common security attacks is known as Primary User Emulation (PUE), where a malicious user tries to imitate a primary user (PU) transmission by emitting wireless signals whose power and waveform characteristics are almost similar to a legitimate PU. In this way, the secondary users (SUs) believe that a PU is present, they vacate the frequency band they were occupying and refrain from using the spectrum. In other cases, adversarial users/devices may attempt denial-of-service (DOS) attacks by jamming the transmissions; poisoning attacks by transmitting corrupted data to manipulate sensing results; spoofing attacks by mimicking transmissions from legitimate users at the physical layer to fool signal authentication systems and intrude protected wireless networks.

For all the aforementioned reasons, the EM signal waveform detection and classification has received increased attention over the last decades. Particularly, Radio Frequency (RF) fingerprinting [1], a technique to identify and classify wireless devices, based on their unique radiometric features present in their received analogue signals, has been utilised as an enhancement to wireless network security at the physical layer. These unique features are primarily a result of the wireless channel response (due to multipaths and the influence of the surrounding environment), which is called *channel-fingerprint*; and the imperfect analogue transmitter components (e.g., filters, mixers, oscillators and power amplifiers), which is called *device-fingerprint*. The majority of RF fingerprinting techniques are either *transient-based* (i.e., aim to extract time- and/or frequency-based features from the signal throughout the transition from the turn-off to the turn-on of a transmitter, just before the transmission) or *steady state-based* (i.e., focus on the unique features, such us I/Q imbalance and frequency/amplitude/phase errors, extracted from the modulated part of the signal, usually the preamble). RF fingerprinting has been successfully explored for a number of wireless devices operating on various standards, such as, Bluetooth, Wi-Fi, radio frequency identification (RFID), wireless sensor networks, cellular and FM transmitters, and has demonstrated that it can be utilised to passively identify the source of transmission [2]. However, transient-based techniques are complex, susceptible to noise and require a higher sampling rate to extract the transient signal due to its short period, while steady state-based techniques require some information of the used modulation scheme.

Over the last years, deep learning techniques have been

This work was supported by the Engineering and Physical Sciences Research Council of the UK (EPSRC) Grant number EP/S026657/1, and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing.

increasingly used to characterise and classify wireless communications [3], [4] and radar signals [5]. However, the vast majority of such work has been treating the EM waveform detection and classification as a supervised learning problem, assuming the existence of prior information (i.e., labelled samples) from all devices/waveforms under consideration. In most cases though, unauthorised or malicious transmissions are originated from previously unidentified users/devices with unknown RF fingerprints. As a result, samples from these waveforms are usually of insignificant size to be effectively modelled or even there are no samples since they appear for the first time. The lack of pre-existing labelled data from such devices makes their identification through traditional supervised learning techniques extremely difficult.

In order to identify such signals, we should rely on techniques, which only use the already available data to learn how to detect any irregular or unobserved patterns/features in a new set of data. This task has been known as *anomaly* or *novelty detection*. Therefore, in the remaining of this paper, we will refer to any already known waveforms (and their corresponding features) as "normal", while any previously unseen waveforms, with features different to the former ones, as "anomalies". In this work, we leverage the latest research advances in machine learning to develop a new system for anomaly detection. The system is trained on a set of known, friendly EM waveforms to learn their "normal" latent space distribution. Then identify any unknown waveforms, which do not belong in that space, as (potentially adversarial) "anomalies".

More specificity, the contribution of this paper can be summarised as follows:

- First, a new dual Autoencoder (AE) enhanced Generative Adversarial Network (GAN) architecture for EM waveform anomaly detection is proposed and implemented . It relies only on already known (presumably friendly) data for training, to identify any previously unseen types of EM waveforms during testing, and classify them as "anomalies" (Section III). The proposed system does not require any prior knowledge of the anomalous (and potentially adversary) signals or data.
- Second, in order to avoid instabilities and mode collapse during training, the Jensen-Shannon (JS) divergence (used as the objective function in previous GAN based architectures for anomaly detection) is replaced by the Wasserstein-1 distance between the real and generated data distributions respectively. Moreover, in order to enforce 1-Lipschitz continuity on the critic when calculating the Wassesrtein distance, "gradient penalty" for weight regularization (which requires very little hyper-parameter tuning) is used (Section II).
- Third, the proposed anomaly detection system is successfully trained and tested on synthetic EM radar and wireless communication waveforms, showing that it can be used to effectively detect previously unseen signals and classify them as potentially adversarial (Section V).

The remaining of this paper is organised as follows. In Section II, a brief introduction on GANs and a summary of their use for anomaly detection is given. Our proposed system to detect previously unseen/potentially adversarial EM waveforms, by characterising them as anomalies, is presented in Section III. Section IV describes the experimental set-up and the EM waveform dataset used for training and testing the proposed approach. Performance results and discussion are given in Section V and finally conclusions are drawn in Section VI.

II. GANs for Anomaly Detection

The Generative Adversarial Networks (GAN) framework, introduced by Goodfellow *et al.* [6] has been successfully applied to model complex and high dimensional distribution of real-world data. These characteristics of GANs suggest that they can be successfully used for anomaly detection, although their application has been only recently explored. In their original formation, GANs consist of two competing networks: a generator G that learns how to map samples drawn from an arbitrary random distribution (usually Gaussian or uniform) to the real data space, and a discriminator D that learns to distinguish between the samples generated by G and real data samples. This can be formulated as a min-max game in which the two players (i.e., G and D) compete against each other, as follows:

$$\begin{aligned}\min_G \max_D \; V(D,G) &= \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] \\ &+ \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))].\end{aligned} \tag{1}$$

After successful training, the discriminator D will learn the real data distribution and should be able to distinguish between real samples and any other set of data samples (including data that have not been encountered before), which do not follow the same distribution. In other words, the probability output of the discriminator can be directly used as an anomaly score. One of the first GAN-based systems to identify rogue transmitters by exploiting this property was proposed in [7]. The objective of the system was to learn hardware impairments in commercial off-the-shelf (COTS) wireless transceivers, such as the in-phase (I) and quadrature (Q) components of the transmitted signal, commonly known as the "I/Q imbalance", which is caused by imperfections in local oscillators and mixers and its unique to different radio hardware. The model was trained on raw I/Q signal data, collected from a number of universal software radio peripheral (USRP) radios. It was shown that the proposed system can accurately identify "trusted" transmitters at high SNR (30dB).

However, as we explain in the following, only the discriminative benefit of the network was exploited in this case, i.e., to minimize the distance between the real and generated sample distributions, respectively. Moreover, vanilla GANs are susceptible to training instabilities and mode collapse and therefore are different to train.

A. Feature space in GANs

The theoretical basis for the connection between the sample and latent space distributions in GANs was first established in

BiGAN [8] and ALI [9], enabling the accurate learning of feature representations. AnoGAN [10] was the first attempt to utilise the latent feature space of a GAN for anomaly detection. During training in [10], only normal samples were used to learn a latent feature space, which captures the normality underlying the given data, so that the corresponding reconstructed (i.e., mapped from the learned latent space back to the original data space) samples and initial samples are as similar as possible. After training, the difference between the input and the reconstructed samples will indicate anomalies. The main limitation of that work is the computational inefficiency due to the iterative search for the latent vector. To address this issue, EGBAD [11] adds an encoder to map input samples to their latent representation during the adversarial training, which considerably improves the inference speed of the network. GANomaly [12] further improves the generator over the previous work by modifying the generator network to an encoder-decoder-encoder network and adding two more extra loss functions to constrain the latent space. All the aforementioned work was primarily designed for, and tested on, digital image datasets. Our proposed system is mainly influenced by [12] and intends to identify anomalies in EM waveforms.

B. Wasserstein GAN-gp

Vanilla GANs are known to be susceptible to training instabilities (i.e., the discriminator is optimised faster than the generator to the point where the discriminator provides no reliable gradient information and the generator barely learns anything) and mode collapse (i.e., the generator learns to generate samples only from a few modes of the training data distribution but fails to create samples from the remaining modes).

In order to overcome these issues, in our proposed network, instead of using the Jensen-Shannon (JS) divergence (used in traditional GANs), we use the Wasserstein-1 distance [13] between the real and generated data distributions as the objective function. By replacing the sigmoidal activation function in the discriminator, its output is now a scalar score (i.e., a value in a range which can be interpreted as how realistic the generated input samples are) rather than a probability. Since the discriminator is not trained to classify inputs as real/fake, it will be renamed to "critic" to reflect its new role, similar to [13]. Moreover, in order to enforce 1-Lipschitz continuity on the critic when calculating the Wassesrtein distance, we use "gradient penalty" [14] (i.e., the squared difference from norm-1) for weight regularization, which requires very little hyper-parameter tuning compared to the intially proposed gradient clipping [13].

III. System Model

The proposed neural network architecture, comprising a dual autoencoder (AE) enhanced generative adversarial network (GAN), is presented in Fig.1. The GAN consists of two main blocks, the generator and the discriminator (which is referred to as the *critic* C, since Wasserstein loss is used). The generator itself consists of two serially connected AEs. The first AE (G_1) consists of an encoder G_1^E, which takes as input any EM signal $\boldsymbol{x}$ and outputs a low dimension vector $\boldsymbol{z}$ (a latent representation of $\boldsymbol{x}$), and a decoder G_1^D, which receives $\boldsymbol{z}$ and outputs $\hat{\boldsymbol{x}}$ (a reconstructed version of the original signal $\boldsymbol{x}$). The output $\hat{\boldsymbol{x}}$ of G_1 is provided as input to both the second AE (G_2) and the critic C. Similar to G_1, the AE G_2 uses an encoder G_2^E to calculate $\hat{\boldsymbol{z}}$ (the latent representation of the generated signal $\hat{\boldsymbol{x}}$), and a decoder G_2^D to map $\hat{\boldsymbol{z}}$ to $\hat{\hat{\boldsymbol{x}}}$ (a reconstruction of $\hat{\boldsymbol{x}}$). The architectures of G_1 and G_2 are identical.

The remaining block of the network architecture is the critic C which receives as input both the real EM signal $\boldsymbol{x}$ and its reconstruction $\hat{\boldsymbol{x}} = G_1(\boldsymbol{x})$. The objective of the critic is to maximize the Wasserstein distance between the two inputs signals. Since "gradient penalty" is used to enforce 1-Lipschitz continuity, the critic's loss function is given by

$$\begin{aligned}\mathcal{L}_{critic} =& \mathbb{E}_{\boldsymbol{x}\sim p_x}[C(G_1(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{x}\sim p_x}[C(\boldsymbol{x})] \\ &+ \lambda_{gp}\mathbb{E}_{\tilde{\boldsymbol{x}}\sim p_{\tilde{x}}}[(\parallel \nabla_{\tilde{\boldsymbol{x}}} C(\tilde{\boldsymbol{x}}) \parallel_2 -1)^2],\end{aligned} \tag{2}$$

where λ_{gp} is a hyper-parameter that weights the gradient loss (here we set $\lambda_{gp} = 10$) and $\tilde{\boldsymbol{x}}$ is sampled uniformly along a straight line between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$ [14]. Note that $\hat{\boldsymbol{x}}$ is not sampled from a random space as it happens in conventional GANs but it is the generated output of the first autoencoder G_1.

A. Generator Loss

The generator loss is formulated as the weighted sum of three different loss functions, each one of them optimising a different part of the overall architecture. These loss functions are described in the following.

1) Adversarial Loss: The adversarial loss $\mathcal{L}_{adv}$ is related to the Wasserstein discriminator loss (2) and is calculated as the average critic score on generated signals:

$$\mathcal{L}_{adv} = -\mathbb{E}_{\boldsymbol{x}\sim p_x}[C(G_1(\boldsymbol{x}))]. \tag{3}$$

2) Reconstruction Loss: The reconstruction loss $\mathcal{L}_{rec}$ is defined as the weighted sum of all $\boldsymbol{L_1}$ distances between $\boldsymbol{x}$, $\hat{\boldsymbol{x}}$ and $\hat{\hat{\boldsymbol{x}}}$, respectively:

$$\mathcal{L}_{rec} = v_1\mathcal{L}_{rec-1} + v_2\mathcal{L}_{rec-2} + v_3\mathcal{L}_{rec-3}, \tag{4}$$

where, v_1,v_2 and v_3 are the weighting parameters and

$$\mathcal{L}_{rec-1} = \mathbb{E}_{\boldsymbol{x}\sim p_x} \parallel \boldsymbol{x} - G_1(\boldsymbol{x}) \parallel_1 \tag{5a}$$

$$\mathcal{L}_{rec-2} = \mathbb{E}_{\boldsymbol{x}\sim p_x} \parallel \boldsymbol{x} - G_2(G_1(\boldsymbol{x})) \parallel_1 \tag{5b}$$

$$\mathcal{L}_{rec-3} = \mathbb{E}_{\boldsymbol{x}\sim p_x} \parallel G_1(\boldsymbol{x}) - G_2(G_1(\boldsymbol{x})) \parallel_1 . \tag{5c}$$

3) Encoder Features Loss: The encoder features loss $\mathcal{L}_{enc}$, aims to minimize the $\boldsymbol{L_2}$ distance between the latent representations, mapped in the low-dimensional feature space by the encoders G_1^E and G_2^E respectively:

$$\mathcal{L}_{enc} = \mathbb{E}_{\boldsymbol{x}\sim p_x} \parallel G_1^E(\boldsymbol{x}) - G_2^E(G_1(\boldsymbol{x})) \parallel_2 . \tag{6}$$

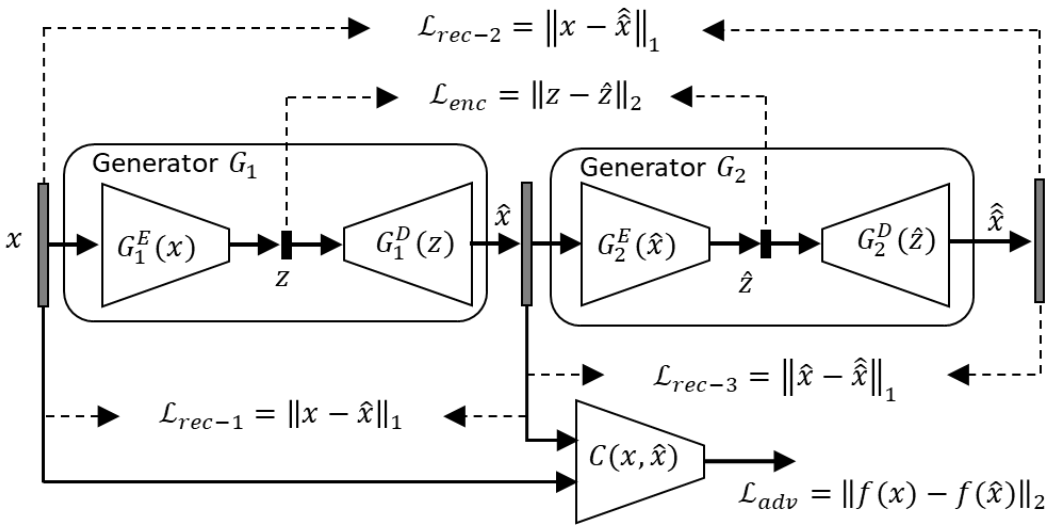

Fig. 1. The proposed dual-autoencoder enhanced GAN architecture, where $f(.)$ has to be a 1-Lipschitz function.

The model is forced to learn latent feature representations which retain the important information needed to reconstruct the input signals. Since the model is optimized towards normal signal instances only, it will fail to minimize the distance between latent representations mapped in the space by other signals (such as, anomalies).

The objective function of the generator is to minimize the following weighted sum of the aforementioned loss functions

$$\mathcal{L}_{gen} = w_{adv}\mathcal{L}_{adv} + w_{rec}\mathcal{L}_{rec} + w_{enc}\mathcal{L}_{enc}, \tag{7}$$

where w_{adv},w_{rec} and w_{enc} are the weights to adjust the impact of the corresponding losses to the overall objective function.

B. Anomaly Score

While all the aforementioned loss functions have been used to train the network, only the latent space representations, captured by the encoder feature loss (6), are used to calculate the anomaly score. Therefore, the anomaly score for a test sample $\dot{\boldsymbol{x}}$ is given by:

$$A_{score}(\dot{\boldsymbol{x}}) = \| \, G_1^E(\dot{\boldsymbol{x}}) - G_2^E(G_1(\dot{\boldsymbol{x}})) \, \|_1 \, . \tag{8}$$

Higher anomaly score implies higher chance that $\dot{\boldsymbol{x}}$ has been picked from a dataset whose distribution is dissimilar to the original training set. However, since the anomaly score is calculated as the $\boldsymbol{L_1}$ norm, which can take any positive value, and the network has no prior knowledge of the "anomalous" data before testing, it is difficult to define a direct anomaly score threshold which separates normal data from anomalies. This issue will be further discussed in the next section.

In order to make the anomaly score easier to interpret, we calculate the whole set of anomaly scores $\mathcal{S} = \{s_i : A_{score}(\dot{x}_i)\}$ for every sample $\dot{x}_i$ in the test set, and apply feature scale (similar to [12]) to have the anomaly scores within the probability range of $[0, 1]$:

$$s_i' = \frac{s_i - \min(\mathcal{S})}{\max(\mathcal{S}) - \min(\mathcal{S})}. \tag{9}$$

IV. Experimental Set-up and EM Dataset

In our experimental setup, we adopted and modified the MATLAB code from [15] to generate a synthetic dataset of different channel-impaired EM radar (i.e., rectangular, linear frequency modulation (LFM) and Barker code) and wireless communication (i.e., BPSK, QPSK, PAM4, GFSK and CPFSK) modulated waveforms. In this paper, our main focus is on radar waveforms, while communication signals are treated as emissions from adversarial sources. The objective of our network is to classify any samples from the training class as "normal" and samples coming from any other waveform class as "anomalies". More specifically, we consider three different scenarios where one radar modulation type is considered as trusted and is used for training, while the remaining radar and communication waveforms are only used in testing.

1) Only Radar Waveforms: In the first scenario, only the three radar waveforms are considered. Among them, only one modulation type is regarded to be available (and assumed to be trusted) in training. The remaining two radar modulation types are available only for testing (and assumed to be adversarial).

2) Radar and Communication Waveforms: Radar systems normally operate in increasingly congested EM environments, competing with other transmitted sources, such as communication systems. Therefore, in the second scenario, we extend the dataset to include both radar and communication waveforms. Data from only one radar waveform type is considered to be present (and trusted) during training, while the remaining two radar modulation types and all five communication waveforms are treated as adversarial signals available only for testing.

3) Only LFM Radar Waveforms: The objective of the last scenario is to investigate whether the proposed system can differentiate between radar signals of the same modulation type having different waveform parameters. We have particularly chosen to test the LFM waveform since it is one of the most widely used for radar systems operating in Low Probability of Interception (LPI) mode, due to its simplicity and effectiveness [5]. A complex representation of a baseband LFM pulse with amplitude A, chirp bandwidth β, and chirp duration τ is given by

$$s(t) = Ae^{j(2\pi\frac{\beta}{\tau}t^2+\phi)}, \text{ for } 0 < t < \tau. \quad (10)$$

Here we consider four waveform classes that correspond to the combination of the following parameter values, $\beta = \{0.1F_s, 0.2F_s\}$ and $\tau = \{10T_s, 20T_s\}$, where F_s and T_s are the sampling frequency and period, respectively. One of the four waveform classes is considered to be present (and trusted) during training, while the remaining three classes are treated as adversarial signals available only for testing.

For each modulation type, 10,000 frames with sample rate of 100MHz is generated, each of them consists of 1024 complex samples. Every time the network is trained on 80% of the trusted signal frames. In testing we use 10% of the trusted signal sample and an equal number of frames uniformly chosen from the remaining signal types. For instance, if LFM is considered as trusted modulation, we use 1,000 samples from LFM and 500 samples from rectangular and Barker, respectively. All signals are impaired with white Gaussian noise with signal-to-noise ratio (SNR) taking one of the following four values [-18, -12, -6, 0, 6] dB. Moreover, the signals are passed through a Rician multipath fading channel. We assume a delay profile of [0 1.8 3.4] samples with corresponding average path gains of [0 -2 -10] dB, and K-factor equal to 4. We consider both static and moving transmitters. For the latter, the maximum Doppler shift is 4 Hz, which is equivalent to a walking speed at 902 MHz carrier frequency.

V. Results

In this work, we use the Area Under the Receiver Operating Characteristics (AUROC) curve metric to evaluate the performance of the proposed "anomaly" classification system. The ROC curve plots the True Positive Rate (TPR) on the y-axis, against the False Positive Rate (FPR) on the x-axis. Therefore, the area under the ROC curve is a measure of the separability between two classes (in our case "normal" and "anomaly"). An ideal model has AUC near 1. In other words, that model can achieve TPR close to 1, while keeping the FPR close the zero. An AUC close to 0.5 means that the model fails to distinguish between classes.

Table I summarises the AUROC values for the first scenario (described in Sec. IV) when the network is trained on one of the EM radar waveforms at a given SNR value and tested on all three radar waveforms with the same SNR. It can be observed that the proposed system is able to learn features in LFM and Barker code modulated waveforms, resulting to AUROC scores close to 1. This means that we can set an anomaly score threshold that allows us to identify "anomalies" (in this case radar modulations, which have not been considered in training) with almost zero false positives. However, for rectangular shape modulation and for low SNR (e.g., -18dB) the AUROC is reduced to 0.832. Lower AUSOC scores indicate that we can not identify anomalies with high probability without suffering from some false positive decision outcomes. This can be visualised in Fig. 2 where the ROC curves for the three radar waveform are presented for SNR at -18dB. Therefore, the anomaly detection threshold should be carefully decided according to our TPR and FPR requirements. The corresponding histogram of the anomaly scores for both normal and abnormal samples, when trained on LFM waveforms at -18dB SNR, is presented in Fig. 3.

TABLE I
AUROC values for RADAR only EM waveforms

SNR	Pulsed Radar Waveforms		
	LFM	**Rectangular**	**Barker**
-18dB	0.981	0.832	0.999
-12dB	0.986	0.911	0.999
-6dB	0.998	0.915	0.999
0dB	0.999	0.968	0.999
6dB	0.999	0.976	0.999

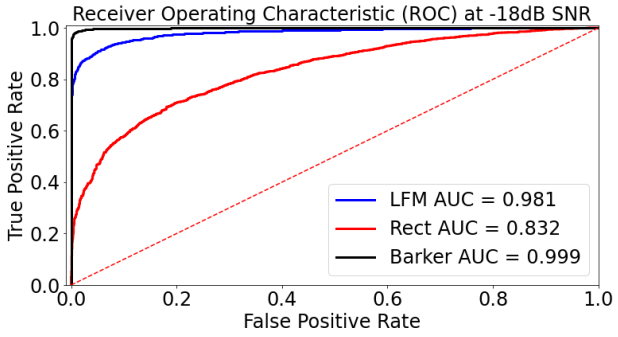

Fig. 2. ROC curves for the 3 EM radar modulated waveforms (SNR=-18dB).

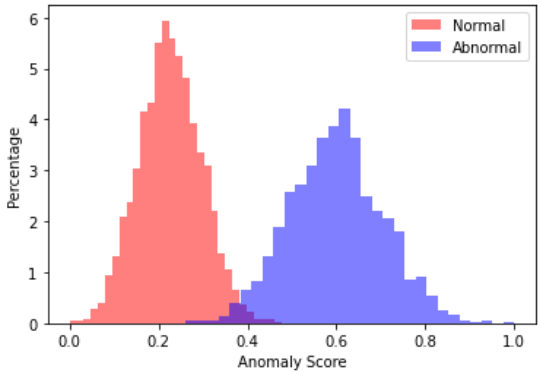

Fig. 3. Histogram of the scores for both normal (i.e., LFM) and abnormal (i.e., rectangular and Barker) test samples (SNR=-18dB).

Table II demonstrates the impact of different SNR values during training and testing on the AUROC for the second

scenario (described in Section IV). More specifically, the system is trained separately on different radar waveforms (i.e., LFM, rectangular and Barker) and SNR values (i.e., SNR={-6, 6}dB), and is tested considering both radar and communication waveforms (i.e., BPSK, QPSK, PAM4, GFSK and CPFSK) of different SNR values (i.e., SNR={-6, 0, 6}dB). Note that the SNR values in training can be different from the testing. It can be seen that different SNR values in training have very small impact on the AUROC for all modulations. However, smaller SNR values during testing seem to have negative impact on the AUROC scores. This is particularly obvious when the system is trained on rectangular pulse modulated waveforms where we have a score reduction of approximately 20% when the SNR is reduced from 6dB to -6dB.

TABLE II
AUROC VALUES FOR RADAR AND COMS EM WAVEFORMS

SNR		Pulsed Radar Waveforms		
Train	Test	**LFM**	**Rectangular**	**Barker**
-6 dB	-6dB	0.999	0.776	0.987
	0dB	0.999	0.921	0.998
	6dB	0.999	0.986	0.989
0 dB	-6dB	0.999	0.871	0.983
	0dB	1.000	0.944	0.999
	6dB	1.000	0.995	0.990
6 dB	-6dB	0.996	0.801	0.974
	0dB	1.000	0.929	0.999
	6dB	1.000	0.991	0.999

Finally, Table III demonstrates the ability of the proposed system to identify different parameters in radar waveforms of the same modulation type. Particularly, in this scenario, we only consider LFM waveforms with different chirp duration τ and bandwidth β, expressed as functions of the sampling period and frequency respectively. The AUROC scores indicate that the proposed system can successfully identify such parameters, particularly at high SNR (i.e., $\geq$0dB). It is worth mentioning that when very similar AUROC scores are observed for different SNR value (e.g., scores very close to 1), the time needed for the system to converge to these scores is much faster for higher SNR values (e.g., 6dB) than for lower ones (e.g., -6dB). This is the case for all three scenarios.

TABLE III
AUROC VALUES FOR LFM ONLY WAVEFORMS

	LFM waveforms			
	$\beta = 0.1F_s$		$\beta = 0.2F_s$	
SNR	$\tau = 10T_s$	$\tau = 20T_s$	$\tau = 10T_s$	$\tau = 20T_s$
-18dB	0.974	0.706	0.974	0.741
-12dB	0.961	0.850	0.999	0.744
-6dB	0.988	0.945	1.000	0.899
0dB	0.996	0.994	1.000	0.999
6dB	0.998	1.000	1.000	1.000

VI. CONCLUSIONS

In this paper, a dual-autoencoder generative adversarial network to detect rogue/hostile radar and wireless communication signal waveforms has been proposed and implemented. After the system is trained on known, friendly EM waveform samples to learn their "normal" latent space distribution, it can identify any unknown samples, which do not belong in that space as (potentially hostile) "anomalies". The proposed system has been successfully trained and tested on a synthetic dataset comprising different pulsed radar and communication modulated signals, impaired with Rician multipath fading and AWGN and it is shown to achieve AUROC scores close to unity even at low SNR. Further studies should investigate and assess the system performance based on real world EM signal waveforms. Finally, an algorithm that exploits the ROC curves to dynamically adapt the anomaly score threshold according to the TPR and FPR requirements, will be part of our future work.

REFERENCES

[1] N. Soltanieh, Y. Norouzi, Y. Yang and N. C. Karmakar, "A Review of Radio Frequency Fingerprinting Techniques," in *IEEE Journal of Radio Frequency Identification,* vol. 4, no. 3, pp. 222-233, Sept. 2020.

[2] S. U. Rehman, K. W. Sowerby, and C. Coghill, "Radio-frequency Fingerprinting for Mitigating Primary User Emulation Attack in Low-end Cognitive Radios," in *IET Communications*, vol. 8, no. 8, pp. 1274-1284, May 2014.

[3] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air Deep Learning Based Radio Signal Classification," in *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.

[4] Y. Shi, K. Davaslioglu and Y. E. Sagduyu, "Generative Adversarial Network in the Air: Deep Adversarial Learning for Wireless Signal Spoofing," in *IEEE Transactions on Cognitive Communications and Networking,* vol. 7, no. 1, pp. 294-303, March 2021.

[5] B. Willetts, M. Ritchie and H. Griffiths, "Optimal Time-Frequency Distribution Selection for LPI Radar Pulse Classification," in *IEEE International Radar Conference (RADAR)*, pp. 327-332, 2020.

[6] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in *Proc. of Inter. Conference on Neural Information Processing Systems (NIPS'14)*, pp. 2672-2680, 2014.

[7] D. Roy, T. Mukherjee, M. Chatterjee, E. Blasch and E. Pasiliao, "RFAL: Adversarial Learning for RF Transmitter Identification and Classification," in *IEEE Transactions on Cognitive Communications and Networking,* vol. 6, no. 2, pp. 783-801, June 2020.

[8] J. Donahue, P. Krähenbühl and T. Darrell, "Adversarial Feature Learning," in *Inter. Conference on Learning Representations* 2017.

[9] V. Dumoulin, M.I.D. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro and A. Courville, "Adversarially Learned Inference," in *Inter. Conference on Learning Representations* 2017.

[10] T. Schlegl, P. Seeböck, S.M.Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," in IPMI 2017, Springer, Cham, pp. 146–157.

[11] H. Zenati, C. S. Foo, B. Lecouat, G. Manek and V. R. Chandrasekhar. "Adversarially Learned Anomaly Detectionn", in Proc. of IEEE International Conference on Data Mining (ICDM), pp. 727-736, 2018.

[12] S. Akcay, A. Atapour-Abarghouei, and Toby P. Breckon. "Ganomaly: Semi-supervised Anomaly Detection via Adversarial Training," in *Asian conference on computer vision*, Springer, Cham, 2018.

[13] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proc. of Inter. Conference on Machine Learning, PMLR* 70:214-223, 2017.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," in *Proc. of Inter. Conference on Neural Information Processing Systems (NIPS'17)*, pp. 5769–5779, 2017.

[15] MATHWORKS, Radar and Communications Waveform Classification Using Deep Learning, 2022 (accessed July 12, 2022). [Online]. Available: https://uk.mathworks.com/help/radar/ug/radar-and-communications-waveform-classification-using-deep-learning.html