

Analyzing the Quality of Synthetic Adversarial Cyberattacks

Ulya Sabeel*, Shahram Shah Heydari*, Khalil El-Khatib*, and Khalid Elgazzar**

*Faculty of Business and Information Technology, **Faculty of Engineering and Applied Science

University of Ontario Institute of Technology, Oshawa, Ontario, Canada

ulya.sabeel@ontariotechu.net, (shahram.heydari, khalil.el-khatib, khalid.elgazzar)@ontariotechu.ca

Abstract—Today’s networked systems face significant security challenges due to sophisticated attacks. Several Machine Learning (ML) and Deep Learning (DL) models are employed to combat these diverse attacks. Adversarial attacks, which can evade detection by AI-based intrusion detection systems (IDS) through small alterations to network attack traffic, pose a significant concern. These AI-synthesized adversarial attacks must adhere to network constraints to seem plausible. In this work, we explore the validation criteria for such adversarial attacks and propose a methodology for analyzing their quality. We evaluate adversarial attack samples synthesized by state-of-the-art generative DL models such as Variational autoencoder (VAE), Conditional Variational autoencoder (CVAE), Generative Adversarial Network (GAN) and compare the performance with our CVAE-Adversarial Network (CVAE-AN) model. Results indicate the effectiveness of CVAE-AN in synthesizing realistic adversarial attacks.

Index Terms—Attack quality, Deep Learning, Generative models, Intrusion detection, Realistic adversarial attacks

I. INTRODUCTION

Artificial Intelligence (AI) is revolutionizing the functioning of present-day networks and security systems. Nevertheless, adversarial attacks present a significant danger to intrusion detection systems that use AI. Research shows that an AI-based IDS can easily be evaded by making slight modifications to the original training data which generates an attack unknown to the IDS [1], [2]. The goal of an adversarial attacker is to manipulate an attack instance in such a way that it is deceptively classified as benign. These attacks can exacerbate potential damage to the organizations that employ such IDS for securing their networks.

The network security research community has shown considerable interest in synthesizing adversarial attacks using DL models [3]–[5]. Multiple defense strategies are employed to enhance the robustness of a model. Adversarial training, for instance, is employed to improve the performance of AI-based IDS against adversarial attacks. This method includes generating adversarial attacks and adding these attacks into the training of an AI-based IDS to improve its robustness against such attacks [2].

However, in the cybersecurity domain, the generated adversarial attacks must resemble realistic network attack traffic. The “quality” of adversarial attacks is determined by their ability to emulate the pattern of values of features for original data while preserving the functional network attack traits. High-quality adversarial attacks closely resemble realistic data samples while introducing small variations in feature

values that can evade detection by the IDS. If adversarial perturbations are added to alter the feature values without maintaining the network constraints, the adversarial attack becomes insignificant from a cybersecurity domain viewpoint.

In this particular context, we define a “*polymorphic attack*” as an atypical attack that mutates its characteristics or feature profile continuously to generate different variants of the same attack to bypass a network’s detection systems while maintaining the functional nature of the attack [2], [6]. When using ML/DL techniques to generate polymorphic adversarial network attacks, care must be taken to ensure the feature values generated by AI are consistent because current AI-based attack generation algorithms do not often take features’ correlations into account. For instance, changing the value of “*packet mean interarrival time*” to generate a new attack without changing the “*packet transmission rate*” is not feasible, as the two features are correlated. Similarly, changing the features representing the statistical characteristics of network traffic is not always possible without considering their correlations.

Attack functionality is also another important factor. A Denial-of-Service (DoS) attack must include a sufficiently high volume to cause resource exhaustion at the target. An AI-generated attack that does not take this fact into account, may generate an attack that is insufficiently impactful to cause any substantial consequence. When considering polymorphic adversarial attacks in which the attacker must change the attack features consistently to evade an AI-enabled IDS, one possible solution is to create a valid attack baseline and then compare subsequent generated attacks with this baseline. The argument is that the modifications in attack features must be significant enough to evade a trained IDS, yet statistically close to the original attack to maintain attack feasibility and functionality. Training a DL model with impractical adversarial data can compromise the model since it learns invalid characteristics and can eventually degrade its robustness and generalization capability for real network scenarios [7]. Therefore, further investigation into the synthesis of better-quality adversarial attacks that resemble realistic network traffic is of utmost importance.

This work is mainly focused on generating realistic adversarial evasion attacks for a network intrusion detection system (NIDS) using generative deep learning. We explore several criteria for ensuring the validity of synthesized adversarial network attacks. Our aim through this research is to develop a

framework to assess the quality of adversarial attacks and use it to evaluate our synthesized attacks. The main contribution of this research includes:

- A methodology to investigate polymorphic adversarial attack realism based on several syntactic and statistical techniques.
- A comparative analysis of the quality of adversarial samples synthesized by several state-of-the-art generative DL models.

The rest of the paper is organized as follows. Section II provides the background of adversarial ML/DL and discusses the relevant related work. Section III describes multiple criteria for ensuring attack quality and our proposed methodology for attack quality analysis. Section IV provides details of our experimental settings and presents an evaluation and analysis of our results. Lastly, section V presents the conclusion for this paper and a brief overview of future scope.

II. LITERATURE SURVEY

Adversarial machine learning is an emerging area of research that aims to assess and enhance the resilience of ML/DL models against deceptive behaviors. Adversarial attack examples are intentionally crafted inputs designed to evade detection by an AI-based IDS.

Several current research works focus on generating synthetic network attack traffic using generative DL models such as Generative Adversarial Network (GAN) [3], Variational Autoencoder (VAE) [4], Conditional Variational Autoencoder (CVAE) [5], and their variants. While these studies emphasize enhancing IDS detection through adversarial training, class balancing, and data augmentation techniques, it is important to note that most of them may not be suitable for real network scenarios where the validity of network data is crucial. Furthermore, there is a lack of emphasis on evaluating the quality of generated synthetic attacks for their effectiveness.

Some recent cybersecurity research works focus on the investigation of the validity of adversarial cyberattacks. Merzouk *et al.* [8], [9] provide a comprehensive analysis of adversarial attacks synthesized using various methods such as the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), DeepFool, Jacobian-based Saliency Map Attack (JSMA) and Carlini&Wagner’s attack (C&W). Based on several invalidation criteria, their results confirm that the attacks generated using the above techniques are likely unrealistic. Vitorino *et al.* [7] introduced an Adaptive Perturbation Pattern Method (A2PM) to generate realistic adversarial attacks based on several network domain and class-specific constraints. Their approach enhances the performance of Multilayer Perceptron (MLP) and Random Forest (RF)-based IDS through adversarial training. Although the authors claim that A2PM produces valid adversarial attacks, the manual selection of features and perturbation of feature values adds complexity and cost to the process.

Apruzzese *et al.* [10] provide an elaborate survey for the analysis of state-of-the-art research using adversarial attacks against ML-based IDS. The authors observe that current threat

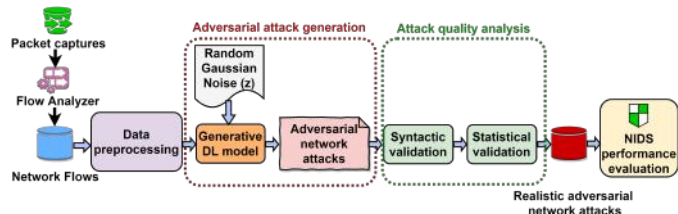


Fig. 1. A generic framework for adversarial attack quality analysis

models are invalid for real network scenarios since they are not entirely black-box systems. To ensure the validity of adversarial attacks, the authors highlight the importance of maintaining the nature of the attack, interdependency among the features, and in-range feature values. Mozo *et al.* [11] employ two Wasserstein GANs to synthesize normal and attack traffic independently enhancing data quality for training a Random Forest (RF) IDS. However, when evaluating the similarity between real and synthetic data using L1 distance and Jaccard Coefficient, the performance of WGAN drops significantly. This indicates the inefficacy of these metrics for adversarial attack quality analysis.

Previous research has explored the validity of adversarial attacks generated using FGSM, BIM, DeepFool, JSMA, and C&W. However, their primary focus is to identify adversarial attacks that are challenging to implement for practical scenarios. In contrast, our research aims to introduce a framework for analyzing the quality of polymorphic adversarial attacks through syntactic and statistical validation techniques.

III. METHODOLOGY FOR ADVERSARIAL ATTACK QUALITY ANALYSIS

A generic framework for adversarial attack quality analysis using our methodology is given in Fig. 1. It consists of four main phases as follows:

A. Network data preprocessing

Network packets are captured and passed through a flow analyzer to extract data flows. These network flows undergo preprocessing which includes cleaning and feature extraction following a standard ML practice. For this work, we adopt the data preprocessing and feature selection phases discussed in our previous research [2], [12].

B. Adversarial attack generation

The preprocessed network attack flows along with random Gaussian noise are fed as inputs to a generative DL model. We employ a Conditional Variational Autoencoder Adversarial Network (CVAE-AN) for this purpose. CVAE-AN is the polymorphic adversarial attack generation and detection model that we introduced in our previous work [2]. The CVAE-AN model builds upon a Semi-supervised GAN (SGAN) by incorporating a CVAE generator in place of a standard Deep Neural Network (DNN) generator of SGAN. During the active phase of an attack, the CVAE attack generator, which is conditioned on the class label produces diverse and evasive adversarial variations of an input attack by changing the feature profile

to successively bypass detection. Further information on the generation of polymorphic adversarial attacks is given in our previous research [2].

C. Adversarial attack quality analysis

Our methodology for analyzing the quality of adversarial attacks comprises two stages: syntactic and statistical validation. Here is a detailed overview of these techniques:

1) *Syntactic validation of adversarial attacks*: The quality of our synthesized adversarial network attacks is first validated using several syntactic constraints such as range coverage for feature values, validity of binary values, and validity of category membership. Range coverage assessment is performed for synthesized features in comparison with original features to make sure they fall within similar ranges to align with network constraints. We scrutinize the validity of binary features which can assume only two values represented by 0 and 1. For instance flags, such as 'Fwd PSH Flags' and other flags in our synthesized dataset are analyzed. This ensures that non-binary values are not erroneously assigned to binary features. Furthermore, the categorical features are examined to check that non-zero values should not be assigned to multiple categories at a time for a given instance. For example, the protocol cannot be set to 1 for both TCP and UDP at the same time.

2) *Kolmogorov-Smirnov hypothesis test*: A two-sample Kolmogorov-Smirnov test [13] is a nonparametric statistical test for comparing the similarity between two datasets. It involves two hypotheses. The null hypothesis H_0 states that the two samples from the two datasets belong to the same distribution. The alternate hypothesis H_1 states that the two samples belong to different distributions.

First, the KS statistic is calculated, measuring the distance between two empirical distributions for all the values of x . Then, the p-value (critical value) is determined, indicating the probability of either rejecting or accepting H_0 . The statistical significance level α is typically set to 0.05. If p-value is equal to or greater than α , H_0 is accepted; if it is less than α , H_1 is accepted.

3) *Hellinger Distance*: Hellinger distance measures the distance between the two probability distributions ranging from 0 to 1 [14]. A value closer to 1 indicates dissimilar distributions, while closer to 0 indicates high similarity.

The Hellinger distance D_h is defined in eq.(1).

$$D_h = \frac{1}{\sqrt{2}} \|\sqrt{P1} - \sqrt{P2}\|_2 \quad (1)$$

Here, $P1$ and $P2$ are the probability distributions for the two data samples respectively.

4) *Correlation Analysis*: Pearson's correlation-based similarity metric [15] measures the pairwise feature correlation between two data distributions indicating their semantic resemblance, with values ranging from 0 to 1. A score of 1 signifies identical pairwise correlations, while 0 implies no resemblance. The similarity between the two correlations is calculated using the eq.(2).

$$Similarity = 1 - \frac{|A_{p,q} - R_{p,q}|}{2} \quad (2)$$

Here, $A_{p,q}$ and $R_{p,q}$ represent the correlation value for the first and second data distribution respectively for a pair of features p and q .

D. NIDS classification and performance evaluation

We identify and choose synthesized adversarial attacks that meet syntactic constraints and closely resemble real network attacks for an in-depth evaluation by our AI-based NIDS. These synthesized flows can also be employed for data augmentation aiding in class balancing and enhancing our system's performance against adversarial attacks.

IV. EXPERIMENTAL SETTINGS AND PERFORMANCE EVALUATION

A. Dataset

We employ the Canadian Institute for Cybersecurity Intrusion Detection System (CICIDS2017) benchmark dataset for this work. This dataset was generated using a variety of realistic network traffic scenarios. It consists of over 80 features with benign records and multiple categories of attack such as DDoS, DoS Slowloris, DoS Slowhttptest, DoS Hulk, DoS GoldenEye, PortScan, Web attack, Bot, SSH-Patator, and FTP-Patator. A more detailed overview of this dataset is given by Sharafaldin *et al.* [16].

B. Generative DL Models and their configuration

For this work, we employ several DL generative models such as VAE, CVAE, and GAN to synthesize polymorphic adversarial attacks using the methodology provided in our previous work on CVAE-AN [2]. The summary of configuration parameters for state-of-the-art generative models such as VAE, CVAE, GAN, and CVAE-AN are sourced from [4], [5], [3], and [2] respectively.

C. Results and discussion

1) *Analysis of results for CVAE-AN*: After syntactic validation of adversarial attacks, we employ several statistical validation techniques such as KS-test, Hellinger distance, and correlation analysis to compare real network attacks with adversarial synthesized polymorphic attacks generated using our model CVAE-AN.

We employ Python's "scipy.stats" library and "ks_2samp" function to measure the KS statistic and p-value for comparing two data distributions. Since the KS test cannot be applied to test all the features in both datasets simultaneously, we compare each feature separately for the real attack and adversarial attack datasets. We then report the average KS-test statistic and average p-value over the entire dataset. Our experiments focus on polymorphic adversarial attacks (Poly AA) generated by CVAE-AN [2], with a specific emphasis on DDoS/DoS attacks due to their severity. However, our methodology can be extended to other attack classes. Table I displays the average KS statistic and p-values for the synthesized polymorphic attacks. Notably, all the adversarial attacks have average p-values equal

TABLE I

AVERAGE KS STATISTIC AND AVERAGE P-VALUES FOR POLYMORPHIC ADVERSARIAL ATTACKS (POLY AA) USING A TWO-SAMPLE KOLMOGOROV SMIRNOV DISTRIBUTION TEST (KS TEST).

Adversarial dataset	Avg. KS Statistic	Avg. p-value
Slowloris DoS Poly AA1	0.52	0.42
Slowloris DoS Poly AA2	0.52	0.42
Slowloris DoS Poly AA3	0.52	0.42
Slow Httptest DoS Poly AA1	0.41	0.50
Slow Httptest DoS Poly AA2	0.44	0.50
Slow Httptest DoS Poly AA3	0.45	0.46
GoldenEye DoS Poly AA1	0.60	0.05
GoldenEye DoS Poly AA2	0.60	0.05
GoldenEye DoS Poly AA3	0.60	0.05
Hulk DoS Poly AA1	0.42	0.42
Hulk DoS Poly AA2	0.42	0.42
Hulk DoS Poly AA3	0.42	0.42
DDoS Poly AA1	0.39	0.44
DDoS Poly AA2	0.39	0.44
DDoS Poly AA3	0.39	0.44

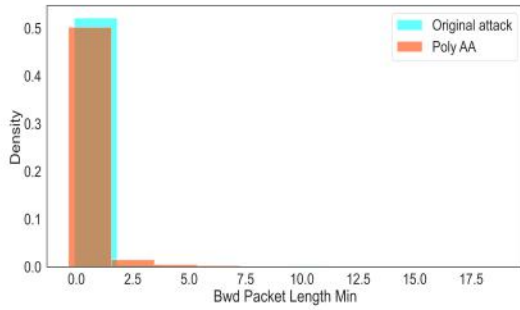


Fig. 2. Slowloris DoS original attack versus polymorphic adversarial attack-1 (poly AA1) for the feature “Bwd Packet Length Min” using Hellinger distance.

to or above the threshold of 0.05, indicating that adversarial attack data follows the same continuous distribution as real attack data.

We apply another statistical distance-based metric to compare the probability distributions of real and adversarially synthesized polymorphic attacks. Similar to the KS test, we compare the distributions of individual features of the two datasets separately. We calculate the distance between them and then find the average distance to determine their similarity. For instance in Fig. 2, we compare Slowloris DoS real attack and Poly AA1 for the feature “Bwd Packet Length Min” using Hellinger distance. The calculated Hellinger distance for this feature is 0.08, indicating a close resemblance to each other. We repeat this process for other attack features and compute the total average distance.

Table II shows the average Hellinger Distance for the adversarially synthesized polymorphic attacks with real attacks. We observe that for most of the adversarial attacks such as Slowloris DoS, Slow Httptest DoS, and Hulk DoS, the Hellinger distance value is closer to 0 indicating their close resemblance to real attacks. The Hellinger distance for attack classes GoldenEye DoS and DDoS, however, lies in the range of 0.31 to 0.52 indicating medium resemblance to the real attacks.

We employ a Pearson correlation-based similarity metric to measure the pairwise feature correlation between real at-

TABLE II

AVERAGE HELLINGER DISTANCE BETWEEN POLYMORPHIC ADVERSARIAL ATTACKS (POLY AA) AND REAL ATTACKS.

Adversarial dataset	Average Hellinger Distance
Slowloris DoS Poly AA1	0.18
Slowloris DoS Poly AA2	0.19
Slowloris DoS Poly AA3	0.18
Slow Httptest DoS Poly AA1	0.20
Slow Httptest DoS Poly AA2	0.17
Slow Httptest DoS Poly AA3	0.30
GoldenEye DoS Poly AA1	0.48
GoldenEye DoS Poly AA2	0.48
GoldenEye DoS Poly AA3	0.48
Hulk DoS Poly AA1	0.14
Hulk DoS Poly AA2	0.15
Hulk DoS Poly AA3	0.15
DDoS Poly AA1	0.52
DDoS Poly AA2	0.43
DDoS Poly AA3	0.31

TABLE III

AVERAGE CORRELATION SIMILARITY SCORE BETWEEN POLYMORPHIC ADVERSARIAL ATTACKS (POLY AA) AND REAL ATTACKS.

Adversarial dataset	Avg. correlation similarity score (in %)
Slowloris DoS Poly AA1	86.08%
Slowloris DoS Poly AA2	86.04%
Slowloris DoS Poly AA3	85.96%
Slow Httptest DoS Poly AA1	78.79%
Slow Httptest DoS Poly AA2	81.73%
Slow Httptest DoS Poly AA3	85.69%
GoldenEye DoS Poly AA1	76.54%
GoldenEye DoS Poly AA2	76.55%
GoldenEye DoS Poly AA3	76.58%
Hulk DoS Poly AA1	84.71%
Hulk DoS Poly AA2	84.29%
Hulk DoS Poly AA3	84.29%
DDoS Poly AA1	80.60%
DDoS Poly AA2	80.77%
DDoS Poly AA3	80.71%

tacks and polymorphic adversarial attacks and analyze the semantic resemblance between the two data distributions. Table III shows the average correlation similarity score for the adversarially synthesized polymorphic attacks with real attacks. We observe that for most of the attack classes, the correlation similarity score of adversarial attack data with real attack data is above 80% indicating that the adversarial polymorphic attacks generated by our system have a close semantic resemblance to real attacks.

The overall results using several statistical validation techniques described previously indicate that our CVAE-AN model can generate polymorphic adversarial network attacks while maintaining the quality of these attacks.

2) *Comparative analysis with state-of-the-art*: We show the effectiveness of our approach in generating better-quality polymorphic adversarial attacks compared to those synthesized by other state-of-the-art DL models such as GAN [3], VAE [4], and CVAE [5]. We select one representative polymorphic adversarial attack from each category for this analysis, but similar results are achieved with other attacks as well.

Fig. 3 provides an assessment of the quality of attacks synthesized using several generative DL models such as CVAE-AN, GAN, VAE, and CVAE. To corroborate our results, we compare the quality of a synthesized polymorphic adversarial attack with an original attack using three tests such as KS test, Hellinger distance, and Correlation similarity. The results for

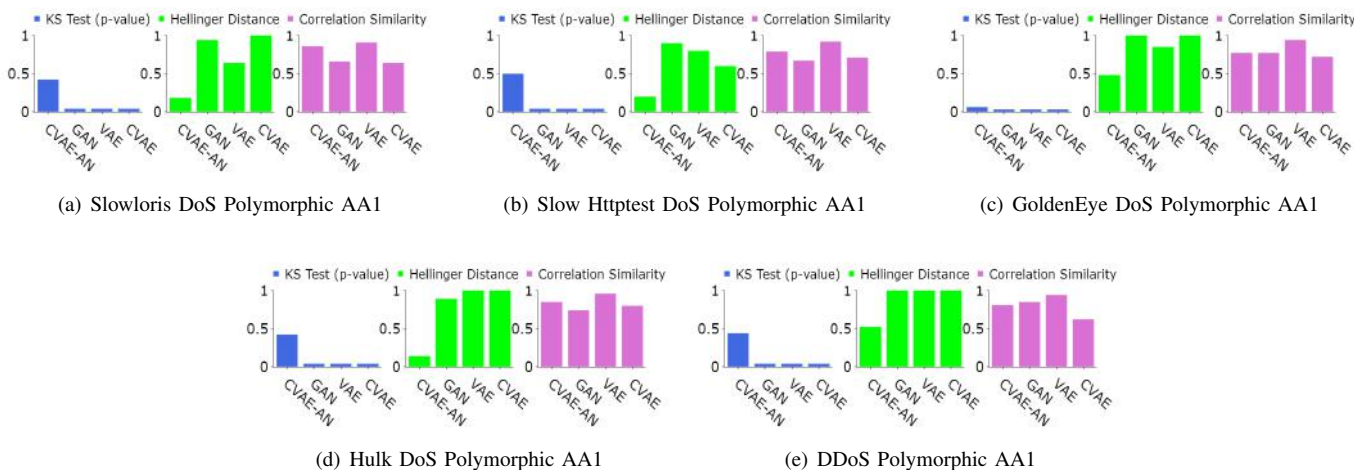


Fig. 3. Polymorphic adversarial attack (AA) quality analysis for state-of-the-art generative DL models.

all the cases of the KS test show the p-value for our CVAE-AN model is the highest (greater than or equal to the threshold value of 0.05) indicating that the attacks generated using our model have a similar distribution as that of the original attacks and therefore are of better quality when compared to attacks synthesized using other DL models.

The Hellinger distance graphs in Fig. 3 indicate that the attacks synthesized using CVAE-AN have the lowest distance from real attacks suggesting their close resemblance to original attacks. Additionally, we observe the correlation similarity of adversarial attacks generated using CVAE-AN to original attacks is higher than most of the other generative models. Overall, results indicate the effectiveness of the CVAE-AN model in synthesizing realistic adversarial attacks.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a methodology for analyzing the quality of AI-synthesized adversarial network attacks. We employ syntactic validation and several statistical techniques to validate attack realism. Moreover, we provide a comparative analysis of polymorphic attacks synthesized by multiple state-of-the-art generative DL models. Our empirical findings suggest that CVAE-AN is the best-performing model when synthesizing realistic polymorphic adversarial attacks.

Acknowledging the importance of addressing adversarial attack risks and enhancing NIDS, we emphasize improving adversarial attack quality. The proposed techniques alone may not guarantee the validity of adversarial attack samples. For future efforts, we aim to focus on specific network constraints for precise and realistic attacks, such as exploring network feature interrelationships for semantic validation.

REFERENCES

- [1] U. Sabeel, S. S. Heydari, H. Mohanka, Y. Bendhaou, K. Elgazzar, and K. El-Khatib, "Evaluation of deep learning in detecting unknown network attacks," in *2019 International Conference on Smart Applications, Communications and Networking (SmartNets)*, 2019, pp. 1–6.
- [2] U. Sabeel, S. S. Heydari, K. Elgazzar, and K. El-Khatib, "Cvae-an: Atypical attack flow detection using incremental adversarial learning," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.
- [3] S. Sapre, K. Islam, and P. Ahmadi, "A comprehensive data sampling analysis applied to the classification of rare iot network intrusion types," in *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2021, pp. 1–2.
- [4] C. Liu, R. Antypenko, I. Sushko, and O. Zakharchenko, "Intrusion detection system after data augmentation schemes based on the vae and cvae," *IEEE Transactions on Reliability*, vol. 71, no. 2, pp. 1000–1010, 2022.
- [5] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot," *Sensors*, vol. 17, no. 9, p. 1967, 2017.
- [6] R. Chauhan, U. Sabeel, A. Izaddoost, and S. Shah Heydari, "Polymorphic adversarial cyberattacks using wgan," *Journal of Cybersecurity and Privacy*, vol. 1, no. 4, pp. 767–792, 2021.
- [7] J. Vitorino, N. Oliveira, and I. Praça, "Adaptive perturbation patterns: realistic adversarial learning for robust intrusion detection," *Future Internet*, vol. 14, no. 4, p. 108, 2022.
- [8] M. A. Merzouk, F. Cuppens, N. Boulahia-Cuppens, and R. Yaich, "A deeper analysis of adversarial examples in intrusion detection," in *Risks and Security of Internet and Systems: 15th International Conference, CRISIS 2020, Paris, France, November 4–6, 2020, Revised Selected Papers 15*. Springer, 2021, pp. 67–84.
- [9] M. Merzouk, F. Cuppens, N. Boulahia-Cuppens, and R. Yaich, "Investigating the practicality of adversarial evasion attacks on network intrusion detection," *Annals of Telecommunications*, pp. 1–13, 2022.
- [10] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, and M. Colajanni, "Modeling realistic adversarial attacks against network intrusion detection systems," *Digital Threats: Research and Practice (DTRAP)*, vol. 3, no. 3, pp. 1–19, 2022.
- [11] A. Mozo, Á. González-Prieto, A. Pastor, S. Gómez-Canaval, and E. Talavera, "Synthetic flow-based cryptomining attack generation through generative adversarial networks," *Scientific reports*, vol. 12, no. 1, p. 2091, 2022.
- [12] U. Sabeel, S. S. Heydari, K. Elgazzar, and K. El-Khatib, "Building an intrusion detection system to detect atypical cyberattack flows," *IEEE Access*, vol. 9, pp. 94 352–94 370, 2021.
- [13] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.
- [14] E. Hellinger, "New foundation of the theory of quadratic forms of infinitely many variable ones," *Journal of pure and applied mathematics*, vol. 1909, no. 136, pp. 210–271, 1909.
- [15] *Synthetic Data Metrics*, DataCebo, Inc., 10 2022, version 0.8.0. [Online]. Available: <https://docs.sdv.dev/sdmetrics/>
- [16] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*. SCITEPRESS, 2018, pp. 108–116.