A Review of VAASI: Crafting Valid and Abnormal Adversarial Samples for Anomaly Detection Systems in Industrial Scenarios

Ángel Luis Perales Gómez*[1], Lorenzo Fernández Maimó[1], Alberto Huertas Celdrán[2] and Félix J. García Clemente[1]
[1] Faculty of Computer Science, University of Murcia, 30100 Murcia, Spain
angelluis.perales@um.es; lfmaimo@um.es; fgarcia@um.es
[2]Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH, CH-8050, Switzerland
huertas@ifi.uzh.ch

***Resumen*—Existing adversarial attacks are not feasible in industrial scenarios since they primarly deals with continuous features and not with categorical features. To enhance cybersecurity in industrial settings, this paper introduces an innovative adversarial attack approach tailored specifically to these environments. This novel technique allows for the creation of targeted adversarial samples valid within supervised cyberattack detection models in industrial scenarios, maintaining consistency of discrete values and correcting cases where adversarial samples appear normal. Validation involved assessing mean error and total adversarial samples generated, comparing against the Projected Gradient Descent method and Carlini & Wagner attack across various parameter configurations. Our proposal achieved the best balance between mean error and generated adversarial samples, demonstrating its superiority.**

***Index Terms*—adversarial attacks, anomaly detection, deep learning, explainable artificial intelligence, industrial systems**

Tipo de contribución: Investigación ya publicada (límite 2 páginas)

I. Introduction

Currently, due to the increase of automation in industrial scenarios, the Anomaly Detection (AD) paradigm is being explored to safeguard devices and technologies against industrial cyberattacks. In particular, AD systems implemented by means of Machine Learning (ML) and Deep Learning (DL) techniques have proven effective in this context. However, this techniques are vulnerable to adversarial attacks, which creates samples that are missclassified by the AD. In this context, we emphasize two significant limitations of adversarial attacks. Firstly, these attacks are ineffective with categorical data commonly generated by industrial devices. Secondly, the existing techniques introduce large errors since they modify the whole set of features. More efforts are required to enhance the robustness of AD and develop suitable adversarial attacks in industrial scenarios.

This paper reviews [1] and introduces VAASI as a solution to the previous limitations. VAASI is a targeted adversarial attack specifically designed for industrial systems, generating valid abnormal samples with minimal error in such environments. Validation was performed using the WADI dataset containing both categorical and continuous features, comparing results with the Projected Gradient Descent (PGD) and Carlini & Wagner (CW) adversarial attacks. Our attack demonstrated the optimal balance between the number of generated adversarial samples and the crafting error, resulting in samples challenging for experts to detect.

II. Implementation of VAASI attack

In this section, we summarized the steps to implement the VAASI attack. In particular, the implementation is divided into five steps: 1) selecting features for modification, 2) generating continuous features, 3) generating categorical features, 4) ensuring the validity of samples, and 5) validation.

Selecting Features for Modification. We suggest modifying particular features to introduce the minimal necessary error and thus create adversarial samples closely resembling the original ones. To identify the appropriate features, we utilize SHapley Additive exPlanations (SHAP) to assess feature importance. After obtaining the importances for each feature, they are arranged in descending order. This enables us to select the top values based on a percentile statistic determined through experimentation.

Generating Continuous Features. In this step, we use PGD to generate continuous features. This attack calculates in each iteration the gradient of the loss function with respect to the input and modulates the gradient sign by a perturbation parameter, to finally subtract it from the original sample.

Generating Categorical Features. This study introduces a novel approach for generating categorical features by replacing selected categorical feature values in each sample with values from the most similar sample in the training dataset.

Ensuring the Validity of the Samples. In this work, we propose extracting rules from Decision Trees (DT) within a Random Forest (RF) to verify if generated samples, misclassified as normal by the AD system, still exhibit anomalous behavior. If the sample is classified as normal by the RF, the adversarial sample is considered to have been transformed into a normal sample during the attack. Subsequently, all paths originating from an abnormal leaf node of all DTs are extracted, and the necessary corrections are calculated to satisfy the conditions for each path leading to an anomalous leaf node in each DT. The path with the lowest error is chosen to modify the sample accordingly.

Validation. Mainly, the validation focuses on two key aspects of adversarial attacks. First, it quantifies the number of adversarial samples that the attack has managed to generate, which helps us understand how easily adversarial

ISBN:978-84-09-62140-8

Tabla I
PERFORMANCE COMPARISON BETWEEN STATE-OF-THE-ART ADVERSARIAL ATTACKS (PGD AND CW) AND OUR PROPOSAL

		PGD(0.1)	PGD(0.3)	PGD(0.5)	PGD(0.7)	PGD(1.0)	CW	Ours
L_1	**Samples generated**	2.8 %	2.8 %	2.8 %	2.9 %	3.1 %	-	**5.4 %**
	Average error	0.005	0.016	0.027	0.038	0.054	-	0.006
L_2	**Samples generated**	3.35 %	9.25 %	22 %	33.4 %	49.5 %	86.25 %	62.25 %
	Average error	0.087	0.248	0.389	0.519	0.721	0.823	**0.028**
L_∞	**Samples generated**	66.80 %	86.10 %	87.30 %	89.75 %	92 %	23.34 %	85.80 %
	Average error	1.844	5.435	9.016	12.540	17.574	0.027	**0.205**

samples are generated. Second, it evaluates the mean error of these samples in relation to the original samples from which they were generated. This error determines which adversarial attacks generate samples that closely resemble the originals, making them more challenging for experts to identify.

III. EXPERIMENTS

This section details the steps to deploy our attack in a real industrial scenario of water distribution. In particular, we launched VAASI against the samples contained in the WADI dataset. To deploy the experiment, a series of requisites were required. In particular, a slight preprocessing of the WADI dataset was carried out, as well as the training of the supervised model that offers sufficiently high performance to be deployed in real environments.

Selecting Features for Modification. In this step, we extracted the importance of each feature in test samples using the SHAP library with a background dataset composed of 100 normal and 100 abnormal samples. Next, we selected from the test dataset the features whose SHAP values were the highest for the abnormal class, i.e., those features whose SHAP values were above the 90th percentile, and grouped them into continuous and categorical features.

Generating Continuous Features. In this step we launched the PGD attack provided by the Adversarial Robustness Toolbox (ART). First, we selected all attack samples in the test dataset. After launching the attack, the majority of the samples (83.62 %) were classified as normal by the AD system, and the remaining samples (16.38 %) were still classified as anomalous.

Generating Categorical Features. We employed the nearest-neighbor algorithm to select similar samples. In particular, to facilitate the nearest-neighbor strategy, we used the KBinsDiscretizer class of the scikit-learn library. This class discretized the training dataset and the previously generated adversarial samples using 10 bins and a uniform discretization strategy. Next, we selected the most similar sample in the training dataset and copied their categorical features into the corresponding features of the adversarial sample.

Ensuring the Validity of the Samples. In this step, we trained an RF using the training dataset to determine if the adversarial attacks retain its abnormal behavior. The RF was trained using the scikit-learn library with the number of estimators set to 10 and the maximum depth of the trees set to 10. The trained RF achieved a 0.969 of F1-score. Subsequently, for all the samples classified as normal, we utilized the decision paths of the DTs trained within the RF to identify the required changes in each sample to revert them to their anomalous behavior.

Validation. In this final step, we compared our results with those obtained using the raw PGD and CW methods. We considered the number of generated adversarial samples and their average error. For PGD, we tested different configurations of the maximum permissible disturbance, ε and we set the iterations and disturbance allowed in each iteration, ε_{step}, to 50 and 0,05, respectively. Regarding CW, we set the iterations and the confidence level to 10 and 0, respectively. Finally, we evaluated our proposal and previous methods using several l_1, l_2, and l_∞ norms. The results are shown in Tabla I, indicating that our proposal achieved the best trade-off between the number of adversarial samples generated and the resulting error. To be specific, for l_1 our solution generated 5.4 % adversarial samples with an error of 0.006. Regarding l_1, our solution generated 62.25 % adversarial samples with an error of 0.028. Finally, for l_inf, our solution generated 85.80 % adversarial samples with an error of 0.205. The difference in the average error between our solution and the existing method is explained by the fact that we only modified the most important features.

IV. CONCLUSIONS

In this study, we introduce VAASI, a novel targeted adversarial attack designed for industrial settings. The novel contributions of VAASI is that the generated adversarial samples are valid and retain their abnormal behavior. Validation using the WADI dataset from a water distribution industrial plant compares VAASI with PGD and CW methods, highlighting its superior balance between the number of generated adversarial samples and the incurred error. Specifically, under different norms, VAASI produced smaller errors compared to PGD and CW, demonstrating its effectiveness in generating valid adversarial samples for industrial systems.

ACKNOWLEDGEMENTS

This work has been funded under Grant TED2021-129300B-I00, by MCIN/AEI/10.13039/501100011033, NextGenerationEU/PRTR, UE, Grant PID2021-122466OB-I00, by MCIN/AEI/10.13039/501100011033/FEDER, UE, by the strategic project CDL-TALENTUM/DEFENDER from the Spanish National Institute of Cybersecurity (INCIBE), by the Recovery, Transformation and Resilience Plan, Next Generation EU, by the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberForce (CYD-C-2020003), and by the University of Zurich (UZH).

REFERENCIAS

[1] Perales Gómez, Á. L., Fernández Maimó, L., Huertas Celdrán, A., & García Clemente, F. J. (2023): "VAASI: Crafting valid and abnormal adversarial samples for anomaly detection systems in industrial scenarios", en *Journal of Information Security and Applications*, vol. 79, 103647, 2023.