2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) | 979-8-3315-1088-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICHORA65333.2025.11017129

Evaluation of Anomaly Detection and Adversarial Attacks in the IEEE 9-Bus System

Emrecan Bingöl
Distance Education Application and Research Center
Karadeniz Technical University
Trabzon, Türkiye
emrecanbingol@ktu.edu.tr

Mustafa Şinasi Ayas
Electrical-Electronics Engineering
Karadeniz Technical University
Trabzon, Türkiye
msayas@ktu.edu.tr

İsmail Hakkı Altaş
Electrical and Electronics Engineering
OSTİM Technical University
Ankara, Türkiye
ismailhakki.altas@ostimteknik.edu.tr

***Abstract*—This study focuses on anomaly detection and the impact of adversarial attacks in microgrids, which are critical components of modern power systems. The performance of a Multilayer Perceptron (MLP) based anomaly detection model was analysed under various operating scenarios and against adversarial attacks, specifically the Fast Gradient Sign Method (FGSM) and Momentum Iterative Fast Gradient Sign Method (MI-FGSM). The performance of the anomaly detection model against adversarial attacks was examined under white-box, grey-box, and black-box scenarios with different perturbation magnitudes. It was evaluated using accuracy, precision, recall, and F1-Score metrics. The results demonstrate that white-box attacks significantly reduce the model's performance, with MI-FGSM being more effective than FGSM. Grey-box attacks had a lower impact on the model's performance compared to white-box attacks. On the other hand, black-box attacks had only a limited effect. Moreover, an increase in perturbation magnitude was found to lead to a significant reduction in the model's performance. These findings emphasize the weaknesses of MLP-based anomaly detection systems in microgrids against adversarial perturbations.**

Keywords—anomaly detection, adversarial attack, adversarial machine learning, false data injection, fast gradient sign method, momentum iterative fast gradient sign method

I. Introduction

Power systems, as one of the critical infrastructures, are undergoing transformation due to increasing energy demand, the growing awareness of clean energy, and the expanding use of information and communication technologies (ICTs) [1]. With the integration of ICTs, power systems, like many physical systems, have found their place in cyberspace and have evolved into cyber-physical systems. This transformation has introduced the concept of Cyber-Physical Power Systems into our lives. Electricity generation plants owned by public or private entities, electricity generation plant operators, transmission networks, transmission network operators, distribution networks, distribution network operators, electricity consumers, and information technology providers in the electric energy sector have all established their presence in cyberspace [2].

Anomaly Detection Systems (ADS) and Intrusion Detection Systems (IDS) are utilized to detect both cyber and physical attacks targeting power systems. ADS are systems designed to identify data and events that disrupt general data patterns. In power systems, anomaly detection is employed not only for cybersecurity purposes but also for various other objectives, such as identifying abnormal grid behaviours, overloads, and unusual consumption patterns [3]. On the other hand, IDS detects cyberattacks on networks and identifies anomalies occurring within them. Intrusion detection is achieved by monitoring irregular behaviours and misuse in the networks [4].

Nowadays, machine learning and deep learning methods are effectively utilized in anomaly detection systems. Sayge et al. [5] utilized an MLP-based false data injection attack (FDIA) detection model for the IEEE 14-bus system. They evaluated the performance of the FDIA detection model against the white-box Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) and Jacobian-based Saliency Map Attack (JSMA). They show that adversarial attacks can significantly reduce MLP-based detection performance. Chen et al. [6] propose an adversarial sample generation method inspired by FGSM. This method is applied to power quality assessment data to evaluate the performance of a Feed-Forward Neural Network (FFNN) against adversarial samples. Similarly, it is applied to short-term building load forecasting data to evaluate a Recurrent Neural Network (RNN) against adversarial samples. The proposed adversarial sample generation method has successfully manipulated both FFNN and RNN models, leading to a decline in model performance and an increase in prediction errors. Tian et al. [7] proposed an approach called the Adaptive Normalized Attack (ANA), which is inspired by DeepFool and produces less perturbation compared to FGSM while inducing a higher misclassification rate. ANA was applied to power quality assessment data and tested using an FFNN model. In addition, adversarial training (AT) was implemented to enhance the model's resilience against adversarial attacks, thereby significantly improving its robustness. Zhang et al. [8] utilized a hybrid model combining convolutional neural networks (CNN) and long short-term memory (LSTM) networks for power quality disturbance data classification. The study demonstrated that the model's performance degraded under white-box and black-box FGSM and MI-FGSM adversarial attacks. It was shown that white-box attacks were more effective than black-box attacks and that FGSM was less effective than MI-FGSM. To improve model performance, iterative adversarial training (IAT) was proposed. The results confirmed that IAT is a more effective defence method than AT.

In this study, an MLP-based anomaly detection model was assessed to evaluate the security and reliability of microgrid under adversarial attacks. The model was evaluated against adversarial attacks, including White-Box and Grey-Box FGSM, White-Box and Grey-Box MI-FGSM, and Black-Box attacks, with perturbation magnitudes of ϵ=0.05 and ϵ=0.1. Experimental findings reveal that the model exhibits significant vulnerability to adversarial samples.

979-8-3315-1088-6/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: ULAKBIM UASL - KARADENIZ TECHNICAL UNIVERSITY. Downloaded on June 23,2025 at 08:53:40 UTC from IEEE Xplore. Restrictions apply.

II. Case Study

A. Microgrid Model

The grid model presented in this study is a modified version of the IEEE 9-Bus System. The parameters of the model are derived from the test system model published by Opal-RT Technologies [9]. This model consists of three synchronous generators, three transformers, six transmission/distribution lines, and three loads. The parameters of the grid model are presented in Table 1, Table 2, and Table 3.

The microgrid model was simulated in MATLAB/Simulink application. Three-phase voltage sources are used instead of synchronous generators in the grid model. Intelligent Electronic Devices (IED) developed for each bus are used for monitoring and controlling the microgrid model.

TABLE I. Transformer Parameters [9]

	T1	T2	T3
Nom. Prim. Volt. (kV, L-L)	24	18	15.5
Nom. Sec. Volt. (kV L-L)	230	230	230
R1 (pu)	1.00E-10	1.00E-10	1.00E-10
L1 (pu)	2.88E-02	3.13E-02	2.93E-02
R2 (pu)	1.00E-10	1.00E-10	1.00E-10
L2 (pu)	2.88E-02	3.13E-02	2.93E-02
Rm (pu)	5.00E+03	5.00E+03	5.00E+03
Lm (pu)	5.00E+03	5.00E+03	5.00E+03

TABLE II. Line Parameters [9]

	Lines Between Buses					
	4-5	4-6	5-7	6-9	7-8	8-9
Length (km)	89.93	97.336	170.338	179.86	76.176	106.646
R0 (Ω/km)	5.88E-1	9.24E-1	9.94E-1	1.15E+0	5.90E-1	5.90E-1
L0 (H/km)	3.98E-3	3.98E-3	3.98E-3	3.98E-3	3.98E-3	3.98E-3
C0 (F/km)	5.89E-9	4.88E-9	5.41E-9	5.99E-9	5.89E-9	5.90E-9
R1 (Ω/km)	5.88E-2	9.24E-2	9.94E-2	1.15E-1	5.90E-2	5.90E-2
L1 (H/km)	1.33E-3	1.33E-3	1.33E-3	1.33E-3	1.33E-3	1.33E-3
C1 (F/km)	9.81E-9	8.14E-9	9.01E-9	9.98E-9	9.91E-9	9.83E-9

TABLE III. Load Parameters [9]

	Load 5	Load 6	Load 8
Nominal Active Power (MW)	125	90	100
Nominal Reactive Power (MVar)	50	30	35

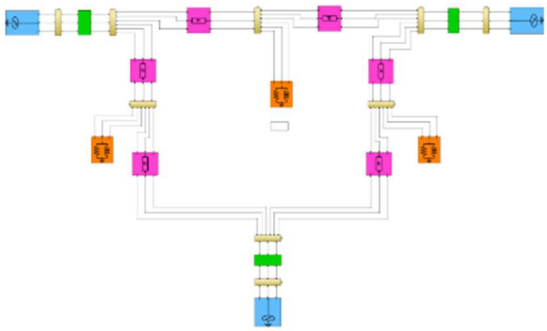


Fig. 1. Simulink Model of Microgrid

B. Proposed IED

IEDs can perform monitoring and protection functions. These devices enable real-time monitoring, fault recording, measurements, and the acquisition of oscillographic data. The use of IEDs adds additional functionalities to electrical grids [10]. PMUs are devices designed to measure the magnitude, phase angle, and frequency deviations of the positive sequence component of voltage and current with high precision and sampling rates. Due to their time-synchronized measurements, PMUs are highly significant devices for Wide-Area Measurement Systems (WAMS). They are used for detecting system anomalies and identifying potential issues [11, 12, 13].

In the microgrid Simulink model, electrical data at the buses is measured by three-phase current-voltage measurement blocks and PMU blocks. The proposed IED includes one three-phase current-voltage measurement block, two PMU blocks, one/two breakers triggered by embedded functions or relays, and mathematical blocks for power measurement.

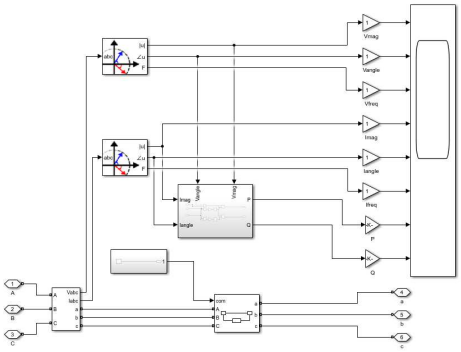


Fig. 2. Simulink Model of Proposed IED

In the three-phase current-voltage measurement block, per-unit three-phase measurement values are transmitted to the PMU, which provides the magnitude, phase angle, and frequency values of the positive sequence components of current and voltage. Active and reactive power values are derived from the magnitude and phase angles.

C. Case Data

The data obtained from each bus consists of eight types: the magnitude of voltage, the phase angle of voltage, the frequency of voltage, the magnitude of current, the phase angle of current, the frequency of current, active power, and reactive power. In a 9-bus model, this results in a total of 72 different data types. The PMU used in the Simulink model collects electrical data by 3.841 measurements per second.

The dataset developed in this study includes forty-one scenarios consisting of normal operation scenarios, fault scenarios, and cyber-attack scenarios. Normal operation scenarios consist of five scenarios simulated in increasing and decreasing caseloads by 5% and 10% in addition to caseload. The fault scenarios include phase-phase, phase-ground, two-phase, two-phase-ground, three-phase, and three-phase-ground faults occurring in the generation and load sections and consist of a total of thirty scenarios. Cyber-attack scenarios are created by manipulating measurements. These manipulations include adding bias values, adding noise, and repeating data. Cyber-attack scenarios consist of six different scenarios.

TABLE IV. Data Classes

Classes	Label	Number Of Data
Normal	0	11737
Fault	1	11520
Attack	2	11136

The normal operation class consists of 11,737 data, the fault class consists of 11,520 data, and the cyber-attack class consists of 11,136 data. In each scenario, duplicate data was cleaned, and then the data was labelled and classified as "0"

Authorized licensed use limited to: ULAKBIM UASL - KARADENIZ TECHNICAL UNIVERSITY. Downloaded on June 23,2025 at 08:53:40 UTC from IEEE Xplore. Restrictions apply.

for the normal operation class, "1" for the fault class, and "2" for the cyber-attack class.

III. Proposed Anomaly Detection and Adversarial Attack Framework

A. Proposed Anomaly Detection System

A Multilayer Perceptron (MLP) is a feedforward Artificial Neural Network (ANN) consisting of an input layer, at least one hidden layer, and an output layer. Feedforward operation refers to the process of transferring input information from the input layer through the hidden layer(s) to the output layer. During the feedforward operation, each neuron computes the weighted sum of the inputs, adds a bias term, and applies an activation function to introduce nonlinear relationships. The loss function, which represents the error between the model's output and the actual value, is calculated. Using this error value, the weights and bias terms are adjusted through backpropagation.

The MLP can adapt to various types of datasets, learn complex relationships, and perform tasks such as classification and regression. These capabilities make MLP widely applicable across many fields. However, key drawbacks include its susceptibility to overfitting, the requirement for large datasets, and substantial computational resources during training.

This study employs a Multi-Layer Perceptron (MLP) for anomaly detection. The model utilizes Rectified Linear Unit (ReLU) activation functions across its four hidden layers. To mitigate overfitting, 30% of the neurons are randomly deactivated (dropout) during each training iteration. The input layer comprises 72 neurons corresponding to the number of features, while the output layer consists of 3 neurons representing the number of output classes, with a Softmax activation function.

B. Proposed Adversarial Attacks

It has been demonstrated that many machine learning methods, including artificial neural networks, are vulnerable to adversarial examples [14]. Adversarial attacks aim to manipulate ML models by adding perturbations to real input data that are imperceptible to humans.

Attacks are generally classified into three types: white-box, grey-box, and black-box attacks. White-box attacks have full knowledge of the target model, including its architecture, parameters, and training data. In contrast, black-box attacks assume no knowledge of the model's internal structure, relying solely on input-output information. Grey-box attacks lie between white-box and black-box attacks, representing a grey area. They are considered a plausible attack model in real-world scenarios [15].

In this study, white-box attacks were conducted using the gradient information of the anomaly detection model. For grey-box attacks, it was assumed that input data and anomaly detection model output were known. Based on this information, a proxy model was trained, and its gradient information was utilized for the attack. In black-box attacks, perturbations with a maximum magnitude of ϵ were added to the input data.

a) FSGM

FGSM is a single-step method proposed by Goodfellow et al. in 2014 [14]. This method generates perturbations proportional to the gradient direction of the target model with respect to the input data, scaled by a factor of ε. The generated perturbations are added to the original input data, causing the target model's outputs to be altered.

$$x_i^{adv} = x_i + \varepsilon \times sign\left(\nabla_{x_i} L(f_\theta(x_i), y_i)\right) \quad (1)$$

In Equation 1, x_i represents the actual inputs, y_i denotes the labels of the input data, $f_\theta(x_i)$ corresponds to the model's output. ∇_{x_i}, is the gradient of the loss function $L(\cdot)$ and $sign(\cdot)$ is the function that computes the direction of the gradient [8].

b) MI-FSGM

MI-FGSM is an iterative method proposed by Dong et al. in 2017 [16]. This method incorporates momentum into the iterative process of the attack, effectively determining the gradient direction and generating efficient perturbations.

$$g_{n+1} = \mu \times g_n + \frac{\nabla_{x_i} L\left(f_\theta\left(x_{i,n}^{adv}\right), y_i\right)}{\left\|\nabla_{x_i} L\left(f_\theta\left(x_{i,n}^{adv}\right), y_i\right)\right\|_1} \quad (2)$$

$$x_{i,n+1}^{adv} = x_{i,n}^{adv} + \frac{\varepsilon}{N} \times sign(g_{n+1}) \quad (3)$$

In Equations 2 and 3, μ represents the momentum decay factor, $\|\cdot\|_1$ denotes the L_1 norm, ε is the perturbation magnitude, ∇_{x_i}, $L(\cdot)$ is the gradient of the loss function and $sign(\cdot)$ s the function that computes the direction of the gradient. Additionally, $g_0 = 0$, $x_{i,0}^{adv} = x_i$ and $x_{i,N}^{adv}$ refers to the adversarial sample generated after N iterations [8].

IV. Experimental Results

The accuracy, precision, recall, and F1-Score values of the MLP model used for anomaly detection are presented in Table IV and V. These metrics illustrate the performance on Test Data (no perturbation), FGSM White-Box Attack, FGSM Grey-Box Attack, MI-FGSM White-Box Attack, MI-FGSM Grey-Box Attack, and Black-Box Attack.

In Table V, the attacks were conducted with $\varepsilon = 0.05$, while in Table VI, the attacks were performed with $\varepsilon = 0.1$. The performance of the anomaly detection model is visualized in Figures 4 and 5.

The evaluation of the MLP-based anomaly detection model under attacks with ε=0.05 revealed that the FGSM White-Box Attack reduced the model's performance by 29.35%-34.04%, FGSM Grey-Box Attack by 17.63%-19.65%, MI-FGSM White-Box Attack by 33.84%-40.54%, MI-FGSM Grey-Box Attack by 20.85%-24.16%, and Black-Box Attack by 1.44%-1.51%.

TABLE V. ADS Performance at Adversarial Attacks

$\varepsilon = 0.05$	Accuracy	Precision	Recall	F1-Score
Test Data	0.9814	0.9823	0.9814	0.9815
FGSM White-Box Attack	0.6934	0.6893	0.6934	0.6474
FGSM Grey-Box Attack	0.8011	0.8091	0.8011	0.7886
MI-FGSM White-Box Attack	0.6493	0.6215	0.6493	0.5836
MI-FGSM Grey-Box Attack	0.7654	0.7775	0.7654	0.7444
Black Box-Attack	0.9673	0.9675	0.9673	0.9674

Authorized licensed use limited to: ULAKBIM UASL - KARADENIZ TECHNICAL UNIVERSITY. Downloaded on June 23,2025 at 08:53:40 UTC from IEEE Xplore. Restrictions apply.

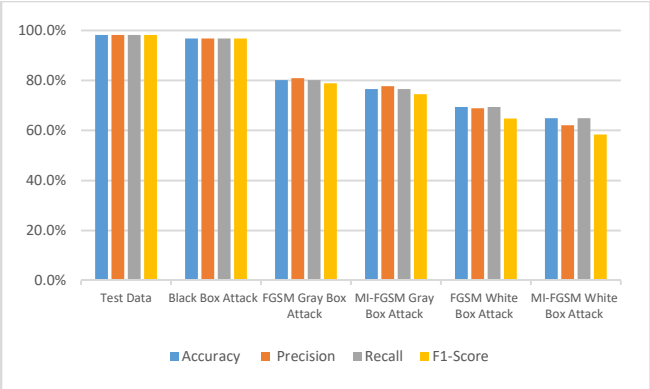


Fig. 3. ADS Performance at Adversarial Attacks (ε=0.05)

Similarly, when evaluated under attacks with $\varepsilon = 0.1$, the model exhibited a performance reduction of 39.57%-49.54% under the FGSM White-Box Attack, 34.66%-40.99% under the FGSM Grey-Box Attack, 56.97%-65.70% under the MI-FGSM White-Box Attack, 43.65%-50.64% under the MI-FGSM Grey-Box Attack, and 3.74%-3.77% under the Black-Box Attack.

TABLE VI. ADS PERFORMANCE AT ADVERSARIAL ATTACKS

$\boldsymbol{\varepsilon = 0.1}$	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Test Data	0.9797	0.9808	0.9797	0.9798
FGSM White-Box Attack	0.5920	0.4949	0.5920	0.4945
FGSM Grey-Box Attack	0.6401	0.6288	0.6401	0.5782
MI-FGSM White-Box Attack	0.4216	0.3135	0.4216	0.3361
MI-FGSM Grey-Box Attack	0.5474	0.5527	0.5474	0.4836
Black-Box Attack	0.9431	0.9438	0.9431	0.9432

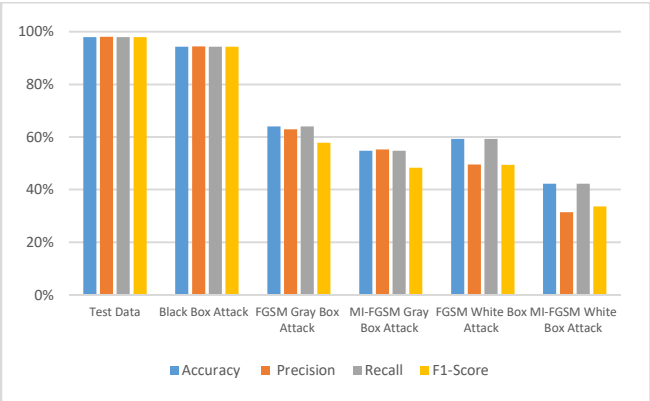


Fig. 4. ADS Performance at Adversarial Attacks (ε=0.1)

This study highlights the vulnerability of MLP-based anomaly detection models in microgrids to adversarial attacks. Experimental results indicate that white-box attacks were more effective than grey-box attacks, whereas black-box attacks had a considerably limited impact. Moreover, the MI-FGSM method led to greater performance degradation compared to the FGSM method. Additionally, increasing the perturbation magnitude resulted in a noticeable decline in the model's performance.

The findings highlight the need for more robust anomaly detection mechanisms to enhance microgrid security. Our future research will focus on developing robust ADS models that can withstand adversarial attacks while improving detection accuracy under various threat scenarios.

REFERENCES

[1] Y. Yoldaş, A. Önen, S. M. Muyeen, A. V. Vasilakos, and İ. Alan, "Enhancing smart grid with microgrids: Challenges and opportunities," *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 205–214, 2017.

[2] H. Aydın, M. A. Barışkan, and A. Çetinkaya, "Evaluation of the security of energy systems under the scope of cybersecurity," *Güvenlik Bilimleri Dergisi*, vol. 10, no. 1, pp. 151–174, 2021.

[3] S. Madabhushi and R. Dewri, "A survey of anomaly detection methods for power grids," *International Journal of Information Security*, vol. 22, no. 6, pp. 1799–1832, 2023.

[4] R. Samrin and D. Vasumathi, "Review on anomaly based network intrusion detection system," *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, Mysuru, India, 2017, pp. 141–147, doi: 10.1109/ICEECCOT.2017.8284655.

[5] A. Sayghe, J. Zhao and C. Konstantinou, "Evasion Attacks with Adversarial Deep Learning Against Power System State Estimation," *2020 IEEE Power & Energy Society General Meeting (PESGM)*, Montreal, QC, Canada, 2020, pp. 1-5, doi: 10.1109/PESGM41954.2020.9281719.

[6] Y. Chen, Y. Tan and D. Deka, "Is Machine Learning in Power Systems Vulnerable?," *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, Aalborg, Denmark, 2018, pp. 1-6, doi: 10.1109/SmartGridComm.2018.8587547.

[7] J. Tian, T. Li, F. Shang, K. Cao, J. Li and M. Ozay, "Adaptive Normalized Attacks for Learning Adversarial Attacks and Defenses in Power Systems," *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, Beijing, China, 2019, pp. 1-6, doi: 10.1109/SmartGridComm.2019.8909713.

[8] L. Zhang, C. Jiang, Z. Chai, and Y. He, "Adversarial attack and training for deep neural network based power quality disturbance classification," *Engineering Applications of Artificial Intelligence*, vol. 127, Art. no. 107245, 2024.

[9] J. P. Bérard, "IEEE 9 Bus System Example," OPAL-RT Technologies, Montreal, Québec, Canada, Version R0, Oct. 23, 2017. Accessed: Dec. 10, 2024. [Online]. Available: https://www2.kios.ucy.ac.cy/testsystems/wp-content/uploads/2020/03/IEEE9_model_documentation_R0.pdf.

[10] K. M. Muttaqi, J. Aghaei, V. Ganapathy, and A. Esmaeel Nezhad, "Control issues of distribution system automation in smart grids," *Renewable and Sustainable Energy Reviews*, vol. 37, pp. 386–398, Sep. 2014, doi: 10.1016/j.rser.2014.05.020.

[11] U. Cali, F. O. Catak, and U. Halden, "Trustworthy cyber-physical power systems using AI: dueling algorithms for PMU anomaly detection and cybersecurity," *Artificial Intelligence Review*, vol. 57, no. 7, p. 183, 2024.

[12] A. G. Phadke and T. Bi, "Phasor measurement units, WAMS, and their applications in protection and control of power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 4, pp. 619–629, 2018.

[13] J. De La Ree, V. Centeno, J. S. Thorp, and A. G. Phadke, "Synchronized phasor measurement applications in power systems," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 20–27, 2010.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint*, arXiv:1412.6572, 2014.

[15] G. Apruzzese and V. S. Subrahmanian, "Mitigating Adversarial Gray-Box Attacks Against Phishing Detectors," in *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 5, pp. 3753-3769, 1 Sept.-Oct. 2023, doi: 10.1109/TDSC.2022.3210029.

[16] Y. Dong, F. Liao, T. Pang, X. Hu, and J. Zhu, "Discovering adversarial examples with momentum," *arXiv preprint*, arXiv:1710.06081, 2017.

Authorized licensed use limited to: ULAKBIM UASL - KARADENIZ TECHNICAL UNIVERSITY. Downloaded on June 23,2025 at 08:53:40 UTC from IEEE Xplore. Restrictions apply.