2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA) | 979-8-3503-4588-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/SERA57763.2023.10197774

Towards an Adversarial Machine Learning Framework in Cyber-Physical Systems

John Mulo*, Pu Tian*, Adamu Hussaini*, Hengshuo Liang*, and Wei Yu*
*Department of Computer and Information Sciences
Towson University, Towson, MD 21252
Emails: {jwere1, ptian1, ahussa7, hliang2}@students.towson.edu, wyu@towson.edu

***Abstract*—The applications of machine learning (ML) in cyber-physical systems (CPS), such as the smart energy grid has increased significantly. While ML technology can be integrated into CPS, the security risk of ML technology has to be considered. In particular, adversarial examples provide inputs to a ML model with intentionally attached perturbations (noise) that could pose the model to make incorrect decisions. Perturbations are expected to be small or marginal so that adversarial examples could be invisible to humans, but can significantly affect the output of ML models. In this paper, we design a taxonomy to provide the problem space for investigating the adversarial example generation techniques based on state-of-the-art literature. We propose a three-dimensional framework containing three dimensions for adversarial attack scenarios (i.e., black-box, white-box, and gray-box), target type, and adversarial examples generation methods (gradient-based, score-based, decision-based, transfer-based, and others). Based on the designed taxonomy, we systematically review the existing research efforts on adversarial ML in representative CPS (i.e., transportation, healthcare, and energy). Furthermore, we provide one case study to demonstrate the impact of adversarial examples of attacks on a smart energy CPS deployment. The results indicate that the accuracy can decrease significantly from 92.62% to 55.42% with a 30% adversarial sample injection. Finally, we discuss potential countermeasures and future research directions for adversarial ML.**

***Keywords*—Adversarial machine learning, Adversarial examples, Cyber-physical systems, Defense, Cybersecurity**

I. Introduction

The applications of machine learning (ML) in cyber-physical systems (CPS) (transportation, healthcare, energy, etc.) have increased significantly [1]–[4]. For example, ML technology has been applied not only to the logistics and supply chain management to reduce waste and inefficiencies, but also to the optimization of energy distribution and consumption, resulting in cost savings and reduced environmental impact [5].

While ML technology has shown great potential in CPS, a variety of attacks on ML technology have recently surfaced. For example, in the training phase, poisoning attacks, which introduce adversarial examples to negatively affect the prediction accuracy of the learning model, might harm the initial probability distribution of training data [6]–[8]. By creating a particular input example without altering the target ML system, evasion attacks can deceive a target system at the testing stage. Additionally, deep learning (DL) has showed promising results in a variety of practical applications, including language translation, speech recognition, cybersecurity, and image captioning, among others. Furthermore, the adoption of ML/DL in different CPS applications, such as traffic management, power grid threat detection, and industrial control system (ICS) anomaly detection, has become increasingly popular in the last decades, driven by the advancements in communication and computational technologies [2].

Nonetheless, adversarial examples [9] can have significant impacts on ML/DL use, such as in energy CPS [10] and in transportation CPS, among in others. Generally speaking, adversarial examples are inputs to ML models with intentionally attached perturbations to lead the model to have incorrect predictions. Perturbations are expected to be marginally small so that adversarial examples could be imperceptible to humans and existing anomaly detection schemes, but can significantly affect the output of the model [9]. For instance, Carlini and Wagner [11] described a targeted attack as one, in which given an input x and a target classification t, we can find an updated input x' (adversarial example) similar to x, while recognized as t.

There are also adversarial examples, in which the input x is misclassified as not one of the top-k target classes. The adversarial examples can also be formed such that the probability of input x being a given class is above or below a certain threshold [12]. As a result, adversarial examples could cause ML models to make incorrect predictions, reducing the accuracy of the model's output. This could have serious consequences in the operations of CPS, including smart grid (demand response, dynamic pricing, etc), smart healthcare (medical diagnosis, etc.), and smart transportation (autonomous driving, etc.) [1]. Adversarial examples pose a security risk of an attack on ML models, leading to an inaccurate decision. Adversarial examples are used for adversarial training as well. They could attack the end-to-end pipelines of prediction-planning of the network in the real-world critical infrastructure systems, often leading to emergencies. As edge computing technology advances, distributed devices have been used to collect data for ML models [13]. However, if some of these components are perturbed, it can fundamentally impact the model during training.

In this paper, we investigate adversarial ML in CPS. Our major contributions in this paper are as follows:

- We design a taxonomy to provide the problem space for systematically investigating the adversarial example generation techniques based on the state-of-the-art literature. Our proposed three-dimensional framework containing three orthogonal dimensions for adversarial attack scenarios (i.e., black-box, white-box, and gray-box), target type, and adversarial examples generation methods (gradient-

979-8-3503-4588-9/23/$31.00 ©2023 IEEE
SERA 2023, May 23-25, 2023, Orlando, USA

based, score-based, decision-based, transfer-based and others).
- Based on the designed taxonomy, we systematically review the existing research efforts on adversarial ML in transportation, healthcare, and energy CPS, respectively. Furthermore, we carry out a case study to show the impact of adversarial ML on energy CPS using a publicly available dataset. We also discuss the defense strategies and the future research directions necessary for securing ML in CPS.

The remainder of this paper is organized as follows. In Section II, we introduce the background of CPS and adversarial examples in DL. In Section III, we design a taxonomy to explore the problem space of adversarial examples generation against ML and introduce the research efforts on adversarial ML against several representative CPS based on our defined taxonomy. In Section IV, we use a publicly available dataset as a case study to demonstrate the impact of adversarial examples on attacks in energy CPS deployments. In Section V, we discuss viable defense strategies and open issues as future research directions. Lastly, we give the final remarks in Section VI.

II. Preliminaries

Cyber Physical System (CPS): The vision of CPS tends to be connecting physical "things" via information communication networking technology and incorporate sensing, computing, control, and networking so that monitoring and control of physical objects can be enabled. The advancements in CPS research could enhance the performance, safety, and reliability of a number of critical infrastructure systems [14]. Likewise, CPS technologies are changing how we interact with one another and the systems around us. CPS has stimulated innovation and competition in transportation, energy, and other industries over the last two decades. Artificial intelligence (AI)/ML and CPS integration, particularly for real-time operation, open new research avenues with significant societal ramifications [1], [2]. For instance, the manufacturing industry is commonly associated with CPS, which is also referred to as smart manufacturing or industry 4.0 [3]. CPS optimizes processes by automating manufacturing and creating a single, decentralized platform for entire factories. As a result, automation in manufacturing reduce labor and material costs and make production time efficient.

Adversarial Examples in ML/DL: They are applied, when hostile input data deceive an ML model so that an incorrect prediction is generated [15]. An adversarial example is a severe weakness in DL systems that cannot be disregarded in security-critical systems such as CPS when DL is integrated to such systems [1], [16]. However, in current research, better justifications are needed for the creation of adversarial examples. The existing efforts have addressed the issue by analyzing the root causes of adversarial examples. The causes could include excessive overfitting or poor regularization of the model, which affects the ability of the learning model to predict unknown data [17]. Nonetheless, by including perturbations into a regularized model, adversarial examples could be mainly used to launch different types of cyber attacks to deceive ML classifiers [18]. Sometimes, protecting CPS against adversarial examples attacks is difficult or even impossible, especially when an adversary has the complete knowledge of the targeted model. For example, according to Zhang and Li [17], there are three characteristics of adversarial examples: transferability, adversarial instability and regularization effect.

III. Taxonomy of Adversarial Example Generation Techniques

In this section, we present the taxonomy of adversarial examples techniques. As shown in Fig. 1, we propose a three-dimensional framework containing orthogonal dimensions. We can classify the adversarial attacks using three concepts, based on the knowledge of the model that the adversary has: white-box [19], black-box [20], and gray-box [21]. Furthermore, since the key objective of adversarial attacks is to deteriorate the performance of classifiers on a specific task, they can further be categorised into evasion attacks, model poisoning, and model extraction attacks [22]. In the figure, the X-axis represents different adversarial attack approaches (i.e., black-box, white-box and gray-box), the Y-axis illustrates the target type, and the Z-axis (gradient-based, score-based, decision-based, transfer-based, and others) represent types of adversarial examples.

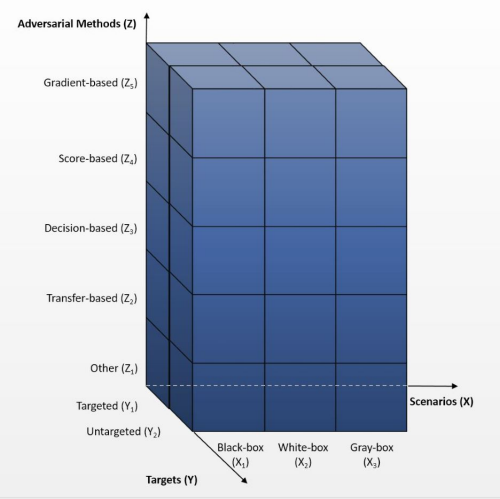

Fig. 1: Problem Space of Adversarial Generation Techniques

A. Adversarial ML Scenarios

1) White-Box: The white-box adversarial ML assumes that the adversary has access to the internal parameters and architecture of the learning model [19]. To initialize a white-box adversarial ML technique, the adversary has to gather the information concerning the architecture and parameters of the model. With the optimal perturbation to the input, the model will output misclassified results. White-box adversarial ML techniques can be very effective since the adversary has the complete knowledge of the architecture and parameters of the model.

2) Black-Box: The black-box adversarial attack technique does not depend on the model parameters or architecture [20]. Still, an adversary can only access the input and output data of the learning model, crafting inputs that causes the model to make incorrect predictions. The black-box adversarial attacks iteratively generate examples, but with perturbations to result in misclassification. The adversary uses the model's gradient parameters regarding the input, which will be used to maximize the prediction error.

3) Gray-Box: A gray-box adversarial attack technique is a method used by an adversary on an ML model that has some, but not complete, knowledge of the internal parameters and architecture of the model [21]. This type of technique lies between the white-box and black-box, in which the adversary has partial information about the model. Gray-box adversarial ML techniques can be more effective than black-box ML techniques as the adversary has some inside information about the architecture and parameters in the model.

B. Target

An adversarial attack can be either targeted or untargeted. On one hand, in a targeted attack, the adversary causes the model to output a particular result or a result in a specific class [12]. On the other hand, an untargeted attack causes the model to output any prediction that is different from the correct one. The adversary's objective in an untargeted attack is to cause misclassification, while not having a preferred forecast for the output. For example, in the context of CPS, misclassification of the industrial component recognition can disrupt the operation of the manufacturing process [23].

C. Adversarial Example Generation Methods

1) Gradient-based Methods: Deep neural networks computes gradients using the error of the desired output and the network output. Gradient-based methods use the gradients in the backpropagation step to produce a perturbation vector on the inputs such that the modified inputs can change the output of the model [24].

For example, FGSM generates adversarial examples by looking for perturbations in finding the way that leads to the loss changes fastest. The CW method finds adversarial examples with small perturbation. Jagielski *et al.* [25] carried a non-targeted adversarial example attack for linear regressions. Adversarial examples generated by ridge regression models can successfully attack neural networks as well. Adversarial examples can also be generated via L-BFGS as improved by Carlini and Wagner and Iterative Target Class Method (ITCM) [26], which enhances FGSM for non-targeted attacks.

2) Score-based Methods: Score-based methods are a kind of black-box techniques used in deep neural networks. These methods perturb the input with an approximation of the gradient obtained from metrics such as probabilities and logits. They include Local search [27], black-box variants of Saliency method and CW [27], as well as ZOO [28], using generative adversarial networks [29], approximate L-BFGS, single pixel, and other similar attacks [30]. Single pixel attacks change single pixels from black to white and vice versa to check the robustness of models. Local search applies the maximum perturbation and checks the accruate ratio of the target class. It then only perturbs the pixels with the highest probabilities of the target class and those around it. Approximate L-BFGS is similar to L-BFGS with the gradient term set to true. The gradients are obtained numerically and the method is suitable for models with small inputs [12].

3) Decision-based Methods: Decision-based methods use only the label output of the model without requiring gradients or probabilities. These methods are more applicable to real-world applications as they only rely on the final output, need no access to the model distribution like transfer-based techniques do, and are more resistant to defenses than gradient-based methods and score-based methods. They include Boundary [30], Spatial [31], Pointwise [32], Hopskip Jump [33], Genattack [34], Precomputed Images, Gaussian Blur, Contrast Reduction, Additive Uniform Noise, etc. [12]. Boundary attack finds perturbations as small as any gradient-based attack by only minimizing the L_2-norm. Pointwise attack is most effective in minimizing the L_0-norm. Additive Uniform Noise performs a line search on perturbations with independent identically distributed (i.i.d.) uniform noise. Precomputed Images uses adversarial images from an external technique similar to the inputs and measure the robustness of the model to the attacks [30]. Line search goes in a random direction from the input and is not that effective due to the large and obvious perturbations as compared to gradient-based, score-based and transfer-based methods.

4) Transfer-based Methods: Transfer-based methods are more difficult as they rely on a substitute model and the training data of the model. Attacks with these methods are more difficult to carry out than decision-based methods that only require the models' output. They rely on the transferability of the adversarial examples from the target model to the victim model. They include FGSM Transfer and Ensemble Transfer techniques [35].

5) Other Methods: Additional techniques include Binarization Refinement, Precomputed Adversarial, and Inversion attacks [36]. To be specific, Binarization Refinement is useful for models that binarize their inputs during data preparation. It uses the binarization data and mapping values to the unperturbed input or over the desired threshold. Precomputed Adversarial uses precomputed adversarial inputs to attack a model. Inversion uses "negative images" by flipping pixel values. These methods do not fit neatly into any of the other categories.

D. Adversarial ML in CPS

Using the designed framework, we now systematically review the representative research efforts on adversarial ML in representative CPS: healthcare, transportation, and energy domain, respectively. Table I shows some examples of research efforts, concerning adversarial examples in transportation CPS using our defined 3-dimensional framework shown in Fig. 1. Likewise, Table II and Table III show some examples of research efforts in energy CPS and healthcare CPS using our defined framework, respectively.

IV. Case Study of Adversarial ML in Energy CPS

This section demonstrates the impact of adversarial ML in energy CPS. The energy CPS critical infrastructure is

TABLE I: Adversarial ML in Transportation CPS

Cube in Framework	References
(X_3, Y_1, Z_5)	[37]
(X_2, Y_1, Z_5)	[38]
(X, Y, Z)	[39]
(X, Y_1, Z_5)	[40]

TABLE II: Adversarial ML in Healthcare CPS

Cube in Framework	References
(X_2, Y_1, Z_2)	[41]
(X_1, Y, Z_2)	[42]
(X_2, Y_1, Z_5)	[43]
(X_2, Y_1, Z_5)	[44]

TABLE III: Adversarial ML in Energy CPS

Cube in Framework	Reference
(X_2, Y_1, Z_1)	[45]
(X_2, Y_1, Z)	[46]
(X, Y_2, Z_5)	[47]
(X_1, Y_1, Z_4)	[48]
(X_1, Y_1, Z_2)	[49]
(X_2, Y_1, Z_5)	[50]

vulnerable to adversarial ML techniques. Grid failure due to an adversarial attack could occur when the ML model is perturbed to produce an unstable condition in the power grid. As a case study, we train a model that takes the input of the factors that lead to a stable grid as well as those that lead to an unstable grid condition. This dataset is based on the UCI machine learning repository, which consists of 10,000 instances on 14 attributes collected on a local stability analysis of the 4-node star system of the decentralized smart grid concept [51].

TABLE IV: Impact of Adversarial Examples in Energy CPS

Adversarial example ratio	Accuracy
0%	92.62%
10%	87.46%
30%	55.42%
50%	42.68%
70%	19.60%

The Decentral Smart Grid Control (DSGC) systems support decentralized power production, such as renewable sources in the smart grid. The fluctuations in renewable energy sources require design and control changes in the power grid to ensure stability and cost efficiency. In this case study, we further show how adversarial examples generated can change the stable and unstable conditions in a model that classifies a DSGC state as stable or unstable.

To simulate the adversarial example generation, we adopt the label flipping as a way of poisoning the training data. In this regard, the model is first trained with clean input with 5,000 examples of grid conditions in stable and unstable states. We use a 4-layer DNN and the accuracy was 92.62% with the clean dataset. As shown in Table IV, the DNN model accuracy rate after training with clean data can vary when false data enforced by label flipping is added.

To be specific, for 10% of the false data that the adversary applies while training, the model's accuracy is 87.46%. The model's accuracy remains much close to the that of the clean model of 92.62%, which indicates some extent of robustness of the model. In addition, a 30% ratio of the false data that the adversary applies in training stage, is whereby the adversary adds false data constituting 30% of the training set labels. The model accuracy dropped significantly to 55.42%. Similarly, for the 50% and 70% scenarios, the accuracy dropped to 42.68% and 19.60%, respectively. According to our experimental results, we observe that adversarial ML can significantly affect the learning model accuracy, which can further impact the operation of energy CPS.

V. Defense Strategies and Future Research Directions

It is common knowledge in the neural network research community that deep learning can be subject to adversarial attacks [1]. Thus, significant research efforts have been devoted on the investigation of adversarial attack methods and defense strategies in ML technology. Adversarial attacks could be classified into several groups based on the target learning model's stages, namely training stage, testing/validation phase, and deployment stage. A number of research efforts in recent literature have proposed defense strategies to deal with adversarial attacks, which can be categorized into data modification, model modification, as well as auxiliary tool-based [52]. We now describe them here briefly.

Data-based Approaches: It refers to altering the input data during the testing or training phases. Some techniques (e.g., adversarial training, blocking transferability, gradient hiding, data compression, and randomization) can be used. Adversarial training can be considered as a meanings of improving the resilience of learning model. In this regard, adversarial samples can be included to the training dataset so that the robustness of the target model can be improved by training the model with correctly labelled adversarial samples [7], [53]. Along with direction, Szegedy *et al.* [53] aimed to make the ML model more resilient against adversaries by modifying the dataset labels and inserting adversarial samples.

Model-based Approaches: It refers to a situation whereby a neural network model can be modified. Examples of methods under this category are deep contractive network, feature squeezing, defensive distillation, regularization, and mask defense. Using deep contractive network as an example, Gu *et al.* [54] proposed a deep compression network employing an autonomous noise reduction encoder, which leads to the reduction of adversarial noise. Their study introduced a smoothing penalty akin to a Convolutional Autoencoder (CAE) during the training phase, which demonstrates to have some defensive effectiveness against attacks (e.g., L-BGFS [53]).

Auxiliary Model-based Approaches: This method uses auxiliary tools (defense-GAN, etc.). For example, Meng *et al.* [55] developed a method known as MagNet, which treats the output of the final layer of the classifier as a black box without altering the classifier or reading any input from the inner layer. In MagNet, a detector was designed to recognize legal and adversarial samples. The framework consists of detector network(s) and a reformer network. In the detector network, a sample will be marked as malicious (rejected) if the distance between the sample and the manifold is larger than the designated threshold. Also, a reformer is designed to convert the adversarial sample into a comparable legal sample based on an automatic encoder. Nonetheless, it is worth noting that as the adversaries could be aware of the MagNet's parameters in white-box attacks, the effectiveness of MagNet could be drastically reduced. To deal with such an issue (i.e., making it difficult for an adversary to obtain the information about which automated encoder is utilized), the hopping strategy of randomly selecting one in a given set of automatic encoders is suggested.

Based on the defense strategies proposed to protect CPS against adversarial examples, the following research directions require further study.

Sophisticated Attacks: The adversary can have intelligence and design more complicated attacking strategies. For example, such attacks are applicable to real-world applications in the wild, and are resistant to defenses such as gradient masking, intrinsic stochasticity, robust training, and defensive distillation [30] as compared to gradient-based and score-based methods. As DL will be used in CPS as a key decision maker, it is critical to systematically evaluate the risk of different ML strategies and understand their impact on the learning model. Additionally, we shall systematically investigate the effect of the learning model's robustness on the performance of CPS. Therefore, to protect the critical DL applications in CPS, more robust learning models shall be designed to deal with different sophisticated attacks.

Our Framework in Other CPS: In this paper, we design a framework for adversarial machine learning in the CPS environment. We study the combination of various scenarios, targets, and adversarial examples generation techniques for adversarial attacks to change either the model or datasets. To demonstrate the security vulnerability, we carry out a case study on the smart grid. We assume that an adversary could change the label collected at distributed nodes. The results show that the accuracy is impacted significantly. Due to the space limitation, we only conduct experiments for energy CPS. We plan to conduct more experiments on additional adversarial ML strategies and other CPS, including smart transportation, smart healthcare, and smart cities, among others. As a future research direction, we will extend our framework to other CPS such as smart cities, smart home, etc.

Adversarial Training as Defense: Adversarial training technique, as a viable defensive strategy, has shown to be effective in [40], [43], [47] against a gradient-based methods in carrying out targeted and untargeted attacks in black-box and a white-box scenarios. However, the trade-off of adversarial training shall be considered as it is computationally expensive in generating adversarial examples for training, especially when its used in CPS with strict performance requirements. Many security related DL applications are vulnerable to attacks, which have been tested against defensive distillation countermeasure. Additional effective attacks are needed to evaluate candidate defensive strategies for high confidence adversarial networks leveraged to confirm the robustness of neural networks trained on different datasets and applications in CPS. For example, defensive distillation could be effective, reducing adversarial attacks success [11]. Distillation is secure against gradient-based attacks (L-BFGS, FGSM, JSMA, etc.).

VI. Final Remarks

In this paper, we addressed the issue of adversarial ML in CPS. In particular, we designed a 3-D framework to categorize the existing adversarial example generation techniques. Our designed framework consists of adversarial attack scenarios, target type, and adversarial examples generation methods. Based on our designed taxonomy, we systematically reviewed the existing research efforts on adversarial ML in transportation, healthcare, and energy CPS. Furthermore, we provided a case study using a smart grid dataset to demonstrate adversarial examples attacks in the CPS deployment. Finally, we discuss the defense strategies proposed to protect CPS against adversarial examples and some future research directions that are required for further study.

References

[1] F. Liang, W. G. Hatcher, W. Liao, W. Gao, and W. Yu, "Machine learning for security and the internet of things: The good, the bad, and the ugly," *IEEE Access*, vol. 7, pp. 158 126–158 147, 2019.

[2] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24 411–24 432, 2018.

[3] H. Xu, W. Yu, D. Griffith, and N. Golmie, "A survey on industrial internet of things: A cyber-physical systems perspective," *IEEE Access*, vol. 6, pp. 78 238–78 259, 2018.

[4] X. Liu, H. Xu, W. Liao, and W. Yu, "Reinforcement learning for cyber-physical systems," in *2019 IEEE International Conference on Industrial Internet (ICII)*. IEEE, 2019, pp. 318–327.

[5] X. Liu, W. Yu, C. Qian, D. Griffith, and N. Golmie, "Integrated simulation platform for internet of vehicles," in *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 2756–2761.

[6] Z. Chen, P. Tian, W. Liao, and W. Yu, "Towards multi-party targeted model poisoning attacks against federated learning systems," *High-Confidence Computing*, vol. 1, no. 1, p. 100002, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667295221000039

[7] C. Qian, W. Yu, C. Lu, D. Griffith, and N. Golmie, "Toward generative adversarial networks for the industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19 147–19 159, 2022.

[8] Z. Chen, P. Tian, W. Liao, and W. Yu, "Zero knowledge clustering based adversarial mitigation in heterogeneous federated learning," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1070–1083, 2021.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[10] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–39, 2022.

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *IEEE symposium on security and privacy (sp)*, may 2017.

[12] J. Rauber, W. Brendel, and M. Bethge, "Foolbox: A python toolbox to benchmark the robustness of machine learning models," in *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. [Online]. Available: http://arxiv.org/abs/1707.04131

[13] F. Liang, W. Yu, X. Liu, D. Griffith, and N. Golmie, "Toward edge-based deep learning in industrial internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4329–4341, 2020.

[14] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, 2017.

[15] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

[16] P. Tian, W. G. Hatcher, W. Liao, W. Yu, and E. Blasch, "Faliotse: Towards federated adversarial learning for iot search engine resiliency," in *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2021, pp. 290–297.

[17] J. Zhang and C. Li, "Adversarial examples: Opportunities and challenges," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2578–2593, 2019.

[18] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[19] Y. Zhang and P. Liang, "Defending against whitebox adversarial attacks via randomized discretization," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 684–693.

[20] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2484–2493.

[21] B. Vivek, K. R. Mopuri, and R. V. Babu, "Gray-box adversarial training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 203–218.

[22] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[23] C. Qian, W. Yu, C. Lu, D. Griffith, and N. Golmie, "Toward generative adversarial networks for the industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19 147–19 159, 2022.

[24] S. Haldar, "Gradient-based adversarial attacks: An introduction," *URL https://medium. com/swlh/gradient-based-adversarial-attacks-an-introduction-526238660dc9*, 2020.

[25] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symposium on Security and Privacy. San Francisco, CA*, may 2018.

[26] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int'l Conf. on Learning Representations*, Toulon, France, 2017.

[27] N. Narodytska and S. P. Kasiviswanathan, "Simple black-box adversarial perturbations for deep networks," *arXiv preprint arXiv:1612.06299*, 2016.

[28] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[29] J. Hayes and G. Danezis, "Learning universal adversarial perturbations with generative models," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 43–49.

[30] W. Brendel, J. Rauber, M. Bethge, and D.-B. Adversarial, "Decision-basedadversarialattacks: Reliableattacksagainstblack-box machine learning models," *Advances in Reliably Evaluating and Improving Adversarial Robustness*, p. 77, 2021.

[31] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "A rotation and a translation suffice: Fooling cnns with simple transformations," *ArXiv*, vol. abs/1712.02779, 2017.

[32] L. Schott, J. Rauber, M. Bethge, and W. Brendel, "Towards the first adversarially robust neural network model on mnist," in *Seventh International Conference on Learning Representations (ICLR 2019)*, 2019, pp. 1–16.

[33] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in *2020 ieee symposium on security and privacy (sp)*. IEEE, 2020, pp. 1277–1294.

[34] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2019, pp. 1111–1119.

[35] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[36] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "On the limitation of convolutional neural networks in recognizing negative images," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 352–358.

[37] A. Omara and B. Kantarci, "Adversarial machine learning-based anticipation of threats against vehicle-to-microgrid services," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 1844–1849.

[38] A. A. D. X. Yulong Cao, Chaowei Xiao and M. Pavone, "Advdo: Realistic adversarial attacks for trajectory prediction," in *European conference on computer vision (ECCV)*. Springer, 2022.

[39] Y. Wang, Z. Su, A. Benslimane, Q. Xu, M. Dai, and R. Li, "A learning-based honeypot game for collaborative defense in uav networks," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 3521–3526.

[40] W. J. Fan LIU, Hao LIU, "Practical adversarial attacks on spatiotemporal traffic forecasting models," in *In Proceedings of the Thirty-sixth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[41] M. Hamidouche, R. Bellafqira, G. Quellec, and G. Coatrieux, "White-box membership attack against machine learning based retinopathy classification," *arXiv preprint arXiv:2206.03584*, 2022.

[42] P. Pandey, M. Chasmai, T. Sur, and B. Lall, "Robust prototypical few-shot organ segmentation with regularized neural-odes," *arXiv preprint arXiv:2208.12428*, 2022.

[43] X. Li, Y. Qiang, C. Li, S. Liu, and D. Zhu, "Saliency guided adversarial training for learning generalizable features with applications to medical imaging classification system," *arXiv preprint arXiv:2209.04326*, 2022.

[44] L. Meng, C. T. Lin, T. P. Jung, and D. Wu, "White-box target attack for eeg-based bci regression problems," *International conference on neural information processing*, pp. 476–488, december 2019.

[45] R. Meyur, A. Pal, M. Youssef, C. L. Barrett, A. Marathe, S. Eubank, A. Vullikanti, V. Centeno, S. Levin, H. V. Poor, *et al.*, "Cascading failures in power grids," *arXiv preprint arXiv:2209.08116*, 2022.

[46] S. Ruj and A. Pal, "Cascading failures in smart grids under random, targeted and adaptive attacks," *arXiv preprint arXiv:2206.12735*, 2022.

[47] B. Manoj, M. Sadeghi, and E. G. Larsson, "Downlink power allocation in massive mimo via deep learning: Adversarial attacks and training," *IEEE Transactions on Cognitive Communications and Networking*, 2022.

[48] N. Kühnapfel, R. Buhren, H. N. Jacob, T. Krachenfels, C. Werling, and J.-P. Seifert, "Em-fault it yourself: Building a replicable emfi setup for desktop and server hardware," in *2022 IEEE Physical Assurance and Inspection of Electronics (PAINE)*. IEEE, 2022, pp. 1–7.

[49] C. Merkel, "Enhancing adversarial attacks on single-layer nvm crossbar-based neural networks with power consumption information," in *2022 IEEE 35th International System-on-Chip Conference (SOCC)*. IEEE, 2022, pp. 1–6.

[50] B. Tu, W.-T. Li, and C. Yuen, "Vulnerability of distributed inverter var control in pv distributed energy system," in *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2022, pp. 47–52.

[51] V. Arzamasov, "Electrical Grid Stability Simulated Data ," UCI Machine Learning Repository, 2018.

[52] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, p. 909, 2019.

[53] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *Proc. Int'l Conf. on Learning Representations. Banff, Canada*, april 2014.

[54] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.

[55] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 135–147.