

A PRACTICAL ADVERSARIAL ATTACK ON CONTINGENCY DETECTION OF SMART ENERGY SYSTEMS

Moein Sabounchi and Jin Wei-Kocsis

Department of Computer and Information Technology - Purdue University

ABSTRACT

Due to the advances in computing and sensing, deep learning (DL) has widely been applied in smart energy systems (SESs). These DL-based solutions have proved their potentials in improving the effectiveness and adaptiveness of the control systems. However, in recent years, increasing evidence shows that DL techniques can be manipulated by adversarial attacks with carefully-crafted perturbations. Adversarial attacks have been studied in computer vision and natural language processing. However, there is very limited work focusing on the adversarial attack deployment and mitigation in energy systems. In this regard, to better prepare the SESs against potential adversarial attacks, we propose an innovative adversarial attack model that can practically compromise dynamical controls of energy system. We also optimize the deployment of the proposed adversarial attack model by employing deep reinforcement learning (RL) techniques. In this paper, we present our first-stage work in this direction. In simulation section, we evaluate the performance of our proposed adversarial attack model using standard IEEE 9-bus system.

Index Terms— Smart Energy System, Adversarial Attack, Procedural Noise, Gabor Noise, Reinforcement Learning, Deep Deterministic Policy Gradient

I. INTRODUCTION

Due to advances in computing and sensing, DL techniques have widely been applied in control management in SESs [1], which has brought upon the advent of emerging tools to effectively address the uncertainties, disturbances, and unforeseen circumstances that may lead to cascading failures of the systems. However, in recent years, increasing evidence shows that DL techniques can be manipulated by adversarial attacks [2]. Adversarial attacks have been introduced on a variety of formats including evasion-attack [3], poisoning attack [4], and inference attack [5]. Based on the knowledge of the attackers, adversarial attacks can also be grouped into white-box, grey-box, and black-box [6]. In recent years, adversarial attacks have been widely studied in computer vision and natural language processing [7], [8], including TextAttack, Fast Gradient Sign Attack (FGSM), projected gradient descent (PGD), and DeepFool attacks [8]–[11]. However, there is very limited work formulating adversarial

attack models deployment and mitigation in SESs. As far as we know, the work in [12] is the only published work in formulating adversarial attack models in SESs. In this work, an effective constrained adversarial attack, called ConAML, is implemented to compromise cyber-physical systems. However, this proposed method mainly focuses on mitigating static operations of energy system. In this paper, we propose an innovative black-box evasion-attack model that is able to practically mislead the decision making of contingency detection in dynamic operation of SESs. To achieve this goal, we develop a practical model to generate additive procedural-noise adversarial perturbation that is used to manipulate the power system measurements and compromise the decision making of contingency detection. To optimize the deployment of generating procedural-noise adversarial perturbation, we design an adversarial RL agent [13] to determine appropriate hyper-parameters of procedural-noise perturbation and maximize the effect of the perturbation on minimizing the accuracy of contingency detection in dynamic control of SESs. Deep RL is a type of DL, where an agent learns how to optimize the actions in an environment by observing the states and seeing the results. Various deep RL techniques have been developed for different application scenarios, such as deep Q-learning [14], advantage actor critic [15], and deep deterministic policy gradient (DDPG) [16]. Due to the nature of our proposed adversarial RL agent that determines optimal hyper-parameters of procedural-noise adversarial perturbation from a continuous space, we decide to employ the DDPG technique.

The rest of this paper is organized as follows, Section II will introduce our proposed black-box evasion-attack for compromising contingency detection in smart energy systems. Sections III and IV will present our performance evaluation and conclusions, respectively.

II. PROPOSED ADVERSARIAL ATTACK MODEL

In this section, we propose an adversarial attack, specifically a black-box evasion-attack, which aims to compromise the detection of contingency in dynamic control of SESs.

A. Threat Model

We formulate the threat model of our proposed evasion-attack as follows: 1) The attacker aims to launch an evasion-attack vector together with other cyber/physical-attack vec-

tors to carry out sophisticated and coordinated attacks; 2) The evasion attack aims to increase the opportunity of coordinated cyber/physical attack to bypass the DL-based contingency detector of the SESs; 3) Because a contingency detector normally has high security priority for dynamic control of SESs, the DL models for contingency detector should be placed with comprehensive and advanced security measures. Therefore, it is reasonable to assume the attack cannot directly access the DL model. Instead, the attack only have limited access to the inputs and outputs of the DL model via eavesdropping and inference attacks [17], [18]. Therefore, the proposed evasion attack is assumed to be black-box; 4) The attacker generates the black-box evasion attack by ensuring certain physical principles, such as voltage stability, for realizing the evasion attack to be stealthy; 5) The attacker can access and compromise a subset of the measurement data that are used by the DL-based contingency detector for inference. This can be realized by man-in-the-middle (MITM) attack [19].

B. Proposed Evasion Attack Model

Based on the threat model stated above, we develop an innovative evasion attack by exploiting procedural noise [20] and RL techniques. The overview of our proposed work is illustrated in Fig. 1. As it can be seen, the SES measurements are used to detect cyber/physical attacks or failures in the system based on a DL-based contingency detector whose behavior can be formulated as a binary/multi-class classifier [21]. The adversarial agent on the other hand aims to mislead the decision making of the DL-based contingency detector. To achieve this goal, the attacker accesses and manipulates a subset of measurement data of the energy system by performing procedural-noise evasion-attack perturbation generation. This process is realized by exploiting Gabor procedural noise technique [22]. The deployment of evasion-attack perturbation generation is optimized by a DDPG-based adversarial RL agent that calculates the optimal hyper-parameters of the Gabor-noise perturbation. This adversarial RL agent generates the hyper-parameters of the procedural noise based on the ground-truth values of the measurement data that are accessible to the attacker, the currently updated Gabor-noise perturbation values, and the decision making of the contingency detector. The procedural-noise evasion-attack perturbation generation mechanism further updates the Gabor-noise evasion-attack perturbation based on the adaptively updated hyper-parameters. To realize stealthy evasion-attack, the generated Gabor-noise perturbation is further constrained based on certain physical principles of the energy system such as voltage stability. By launching the Gabor-noise evasion-attack perturbation that is adaptively optimized by the DDPG-based adversarial RL agent, the compromised measurements (i.e. measurement with additive adversarial noise) are generated to mislead the decision making on DL-based contingency detector and increase the possibility

of coordinated cyber/physical attacks to bypass the detector. Therefore, as shown in Fig. 1, our proposed evasion-attack model mainly consists of two essential components: 1) procedural-noise evasion-attack perturbation generation mechanism that is designed based on Gabor procedural noise technique; and 2) DDPG-based adversarial RL agent for optimizing evasion-attack deployment.

1) Procedural-Noise Evasion Attack Perturbation Generation Mechanism:

One essential challenge of practically carrying out adversarial attacks on dynamic control in energy system is that dynamic control normally requires timely response. Because of this, the gradient-based adversarial attack methods [8]–[11] and the method proposed in [12] can not be directly applied in this application domain. To tackle this challenge, in our work, we exploit procedural noise model to timely generate unified evasion-attack perturbation. The main advantages of procedure noise model includes: 1) requiring low memory and computational complexity, which enables timely generation of adversarial perturbation; 2) being not dependent on a specific measurement data which enables unified adversarial-attack generation; 3) being able to be assessed and optimized independently in an online manner [23]. Perlin [24] and Gabor noise models are two widely adopted procedural noise models. In our work, we exploit Gabor noise model for our evasion-attack perturbation generation.

Procedural noise is non-periodic which can be implemented in N dimensions, where $N > 1$. Procedural noise is parameterized meaning that it can generate a “class of related noise patterns” [23]. The noise functions can generally be classified into three main groups: lattice gradient noises, explicit noises, and sparse convolution noises [23]. For each case, we can present multiple noise functions. As stated previously, we leverage 2-D Gabor noise [22] in our work. Gabor Kernel is the critical component for calculating Gabor noise, which is the multiplication of a “circular Gaussian” function and two-dimensional cosine function:

$$g(x, y) = K e^{-\pi\sigma^2 (x^2+y^2)} \cos(2\pi F_0 (x\cos\omega_0 + y\sin\omega_0)) \quad (1)$$

where K and σ are the magnitude and width of the circular Gaussian function, respectively. F_0 and ω_0 are the frequency and orientation of the cosine function, respectively. Gabor noise can be derived as the weighted sum of Gabor Kernels, which can be described as follows:

$$N(x, y) = \sum_i W_i g(K_i, \sigma_i, F_{0_i}, \omega_{0_i}, x - x_i, y - y_i), \quad (2)$$

where $\{W_i\}$ is the set of weights and $g(\cdot)$ is the Gabor kernel. In our proposed method, we use the absolute values of the measurements in SESs, including voltage, active and reactive power, and frequency, as the x-dimension variable in Eq. (2). Additionally, we use logarithm of the bus indices of

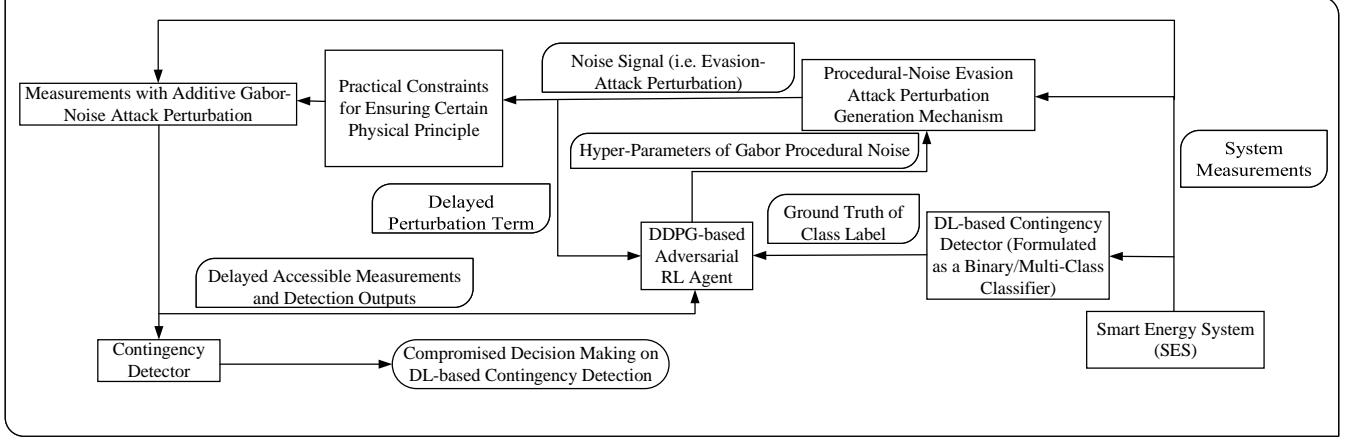


Fig. 1. Illustration of our proposed adversarial attack model.

energy system as the y-dimension variable in Eq. (2), which can be formulated as follows:

$$y_i = \log(i + 1) \quad (3)$$

The authors would like to clarify that there can be alternative functions to $\log(\cdot)$ in generating y_i . For simplicity, we adopt $\log(\cdot)$ in our work. Additionally, as stated above, x_i can be the measurements of voltage, active and reactive power, and frequency. When it is desired to generate Gabor-noise adversarial perturbation associated with more than one types of the measurements, parallel Gabor model needs to be applied.

2) DDPG-based Adversarial RL Agent:

To optimize the deployment of our proposed evasion-attack, we exploit one RL technique, DDPG, to generate appropriate values of the critical hyper-parameters of the Gabor-noise evasion perturbation model, including σ , F_0 , and ω_0 , and maximize the effect of the evasion-attack on minimizing accuracy of detecting contingency on dynamic operation of SESs. To achieve this, we model the procedural-noise perturbation generation for the energy system as a Markov decision process (MDP) that is defined by $(\mathcal{S}, \mathcal{A}, r, \mathbb{P})$. $s_k \in \mathcal{S}$ represents the input states for the system. The action $a_k \in \mathcal{A}$ refers to the decision making on the values of hyper-parameters of the Gabor-noise perturbation model including σ , F_0 , and ω_0 at time $t = k$. The decision making on the action results in the generation of Gabor-noise perturbation n_k . \mathbb{P} is the transition probability that characterizes the dynamics of the SES operation. $r : \mathcal{S} \rightarrow R$ is the reward function for our proposed adversarial RL. The reward function at time $t = k$, is formulated as follows:

$$R(k) = \sum_{\tau=k}^{T_f} \lambda^{\tau-k} R(s_\tau, a_\tau) \quad (4)$$

where $s_k \in \mathcal{S}$, which as mentioned earlier represents the input states at time $t = k$, can be written per the following:

$$s_k = [x_k, n_k, c_k] \quad (5)$$

where x_k is the vector consisting of the absolute values of the measurements associated with the individual buses (e.g. voltage, power, frequency and reactive power vectors individually) at time $t = k$. Specifically, we can describe it as: $x_k = \{x_{k,i}\}$, where i denotes the bus index associated with the individual measurement data in the SES. n_k denotes the Gabor-noise perturbation, c_k denotes the difference between ground-truth and misled outputs of the contingency detector, and λ is the discount factor. Additionally, the reward value at time $t = k$, $R(s_k, a_k)$ in Eq. (4) is formulated as follows:

$$R(s_k, a_k) = c_k - \sum_i e^{k_0(x_{k,i} - \hat{x})} - \sum_i e^{k_0 n_{k,i}} \quad (6)$$

where $c_k, x_{k,i}, n_{k,i} \in s_k$ as defined in Eq. (5), and $n_{k,i}$ is calculated based on a_k , and \hat{x} represents the nominal value for each measurement. In our work, we optimize the structure of the adversarial RL agent via online training, and store the trained agent model for timely usage. The proposed adversarial RL agent is then utilized to adaptively generate appropriate hyper-parameters of the Gabor-noise perturbation in near real time with minimal computational complexity.

III. SIMULATION RESULTS

In this section, we evaluate the performance of our proposed adversarial attack, which is specifically a black-box evasion attack, on compromising the contingency detection in dynamic operation of SESs. We use IEEE 9-Bus System for the performance evaluation [25]. In the simulation, we consider that there is a built-in contingency detector based on a fully-connected deep neural network. Additionally, we assume that the attacker is able to access and manipulate

the voltage measurements, i.e. x_k refers to voltage measurements in our simulations. Furthermore, the attacker aims to develop and carry out the proposed Gabor-noise evasion attack to minimize the accuracy of the contingency detector on detecting a coordinated physical attack on Generator 5 beginning from $t = 5.4s$. Without deploying evasion-attack, the voltage measurements are shown in Fig. 2. In this situation, the posterior probability of contingency occurrence estimated by the DL-based detector is shown in Fig. 3. We can observe that the detector is able to timely detect the fault with high accuracy when our proposed adversarial attack is not launched. Next, we generate our proposed Gabor-noise

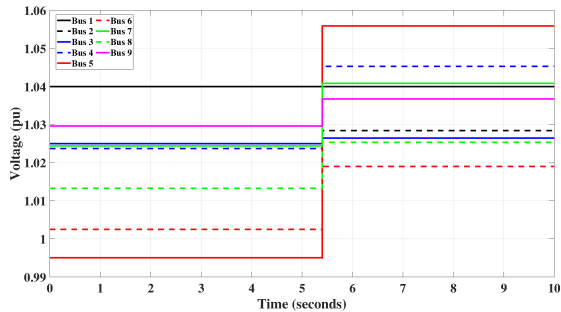


Fig. 2. Voltage measurements with physical fault and without evasion attack.

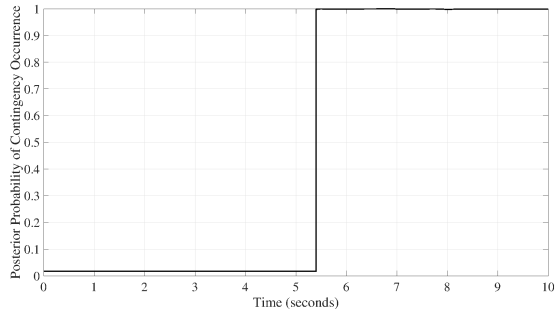


Fig. 3. Performance of the DL-based contingency detector with physical attack.

evasion-attack perturbation whose critical hyper-parameters are optimized by using our proposed DDPG-based adversarial RL agent. The normalized perturbation value is shown in Fig. 4. The perturbation is added to the voltage measurements shown in Fig. 2. The compromised voltage measurements are shown in Fig. 5. From Fig. 5, we can observe that the evasion-attack perturbation only results in a range of voltage-measurement variation within $[0 - 0.01]pu$. In other words, our proposed evasion-attack perturbation is effectively constrained based on a certain physical principle (i.e. voltage stability in our scenario). By using the compromised voltage measurements, the performance of the DL-based detector on detecting the physical attack is shown in Fig. 6. From Fig. 6, we can observe that our proposed evasion-attack can effectively mitigate the performance of the DL-based detector of the SESs.

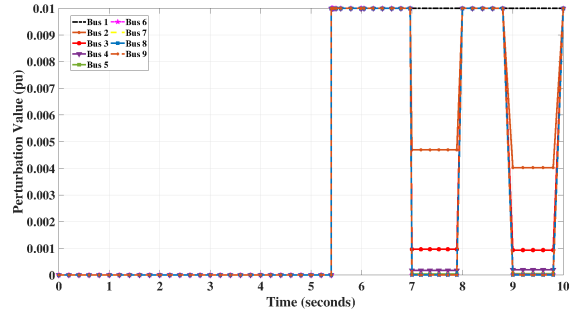


Fig. 4. Values of Gabor-noise evasion-attack perturbation associated with individual buses.

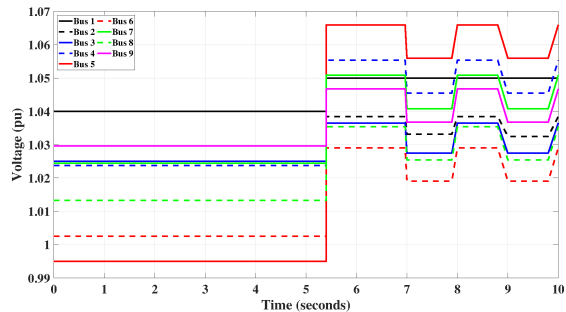


Fig. 5. Compromised voltage measurements by launching our proposed evasion attack.

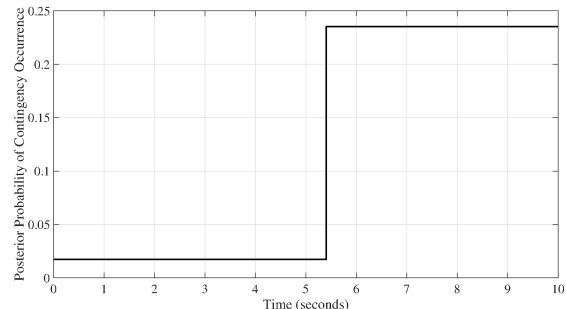


Fig. 6. Performance of DL-based contingency detector with physical attack and evasion attack.

IV. CONCLUSIONS

In this paper, we present our initial-stage work on developing a practical adversarial attacks on dynamic controls of SESs. In this work, we propose an innovative adversarial attack model that is able to practically compromise the contingency detection in SESs. Our proposed model mainly consists of two components: 1) procedural-noise evasion-attack perturbation generation mechanism, and 2) DDPG-based adversarial RL agent for optimizing the evasion-attack deployment process. As illustrated in the simulation results, our proposed evasion attack model can practically compromise the contingency detector in the SESs by resulting in a significant reduction in detection accuracy. In our ongoing

work, we are improving the formulation and realization of the adversarial attack by using larger-scale energy systems, and we aim to further leverage the proposed adversarial attack for enhancing the resilience of the intelligent control mechanism in energy systems.

V. REFERENCES

- [1] Youbiao He, Gihan J. Mendis, and Jin Wei, “Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism,” *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2505–2516, 2017.
- [2] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel, “Adversarial attacks on neural network policies,” *arXiv preprint arXiv:1702.02284*, 2017.
- [3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrđić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, “Evasion attacks against machine learning at test time,” in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.
- [4] Erwin Quiring and Konrad Rieck, “Backdooring and poisoning neural networks with image-scaling attacks,” in *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2020, pp. 41–47.
- [5] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, “Membership inference attacks against machine learning models,” in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.
- [6] Seyed-Mohsen Moosavi-Dezfooli, Ashish Shrivastava, and Oncel Tuzel, “Divide, denoise, and defend against adversarial attacks,” *CoRR*, 2018.
- [7] Naveed Akhtar and Ajmal Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” *IEEE Access*, vol. 6, pp. 14410–14430, 2018.
- [8] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi, “TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, Oct. 2020, pp. 119–126, Association for Computational Linguistics.
- [9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, “Towards deep learning models resistant to adversarial attacks,” 2019.
- [10] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.
- [11] Nicholas Carlini and David A. Wagner, “Towards evaluating the robustness of neural networks,” *CoRR*, 2016.
- [12] Jiangnan Li, Jin Young Lee, Yingyuan Yang, Jinyuan Stella Sun, and Kevin Tomsovic, “ConAML: Constrained adversarial machine learning for cyber-physical systems,” *arXiv preprint arXiv:2003.05631*, 2020.
- [13] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, MIT press, 2018.
- [14] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al., “Deep q-learning from demonstrations,” in *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [15] Mohammad Babaeizadeh, Iuri Frosio, Stephen Tyree, Jason Clemons, and Jan Kautz, “Reinforcement learning through asynchronous advantage actor-critic on a gpu,” *arXiv preprint arXiv:1611.06256*, 2016.
- [16] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.
- [17] TOMM PEAKE, “Eavesdropping in communication,” *Animal communication networks*, p. 13, 2005.
- [18] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang, “Membership inference attack against differentially private deep learning model,” *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.
- [19] Avijit Mallik, “Man-in-the-middle-attack: Understanding in simple words,” *Cyberspace: Jurnal Pendidikan Teknologi Informatika*, vol. 2, no. 2, pp. 109–134, 2019.
- [20] Ares Lagae, Sylvain Lefebvre, George Drettakis, and Philip Dutré, “Procedural noise using sparse gabor convolution,” *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3, pp. 1–10, 2009.
- [21] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou, “Multi-class adaboost,” *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.
- [22] Bruno Galerne, Ares Lagae, Sylvain Lefebvre, and George Drettakis, “Gabor noise by example,” *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 1–9, 2012.
- [23] Ares Lagae, Sylvain Lefebvre, Rob Cook, Tony DeRose, George Drettakis, David S Ebert, John P Lewis, Ken Perlin, and Matthias Zwicker, “A survey of procedural noise functions,” in *Computer Graphics Forum*. Wiley Online Library, 2010, vol. 29, pp. 2579–2600.
- [24] Ken Perlin, “Improving noise,” in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 681–682.
- [25] Vijay Vittal, James D McCalley, Paul M Anderson, and AA Fouad, *Power system control and stability*, John Wiley & Sons, 2019.