A Natural Watermarking Approach to Cyber Attack Detection for Power Electronics-Interfaced Renewables

Imasha Balahewa, Lars Bjorndal, Chris Mi, Tong Huang
Department of Electrical and Computer Engineering
San Diego State University
San Diego, USA
Email: {ibalahewa, lbjorndal, cmi, thuang7}@sdsu.edu

***Abstract*—This paper introduces a novel approach to detecting cyber attacks for power electronics-interfaced renewable resources, e.g., solar panels. The approach leverages the inherent variability of renewable energy generation to watermark the measurements of renewable resources that are vulnerable to false data injection (FDI) attacks. By checking the existence of the watermarks imprinted by the natural fluctuations of renewables, false data injection attacks can be detected. Compared with the conventional watermarking methods, the proposed approach does not require additional noise injection which compromises control performance. The effectiveness of the proposed approach is validated by simulating a solar photovoltaic system.**

***Index Terms*—Inverter-based resources (IBRs), false data injection, cyber security**

I. Introduction

To decarbonize the electricity infrastructure, an increasing number of conventional fossil-fueled generators are being replaced by renewable generation resources in bulk transmission systems, and more distributed energy resources are emerging in distribution grids [1]. Most of these new resources are interfaced with the AC grids via inverters whose dynamics are mainly governed by the control and communication software embedded in the inverters. Due to this coupling between physical and cyber layers, a malicious intrusion into the cyber layers of inverter-based resources (IBRs) can significantly compromise the reliability and efficiency of their host power grids. This has been exemplified by the 2015 and 2016 attacks on Ukraine's power grid which both triggered blackouts affecting over 200,000 customers [2], [3] and the 2010 Stuxnet attack which destroyed up to a fifth of Iran's nuclear centrifuges [4]. It is therefore imperative to detect and mitigate the cyber risks in IBRs in a timely manner [5], [6], [7].

Multiple studies have been conducted to identify and mitigate cyber risks associated with inverter-based resources (IBRs) [8], [9], [10]. Reference [8] presents a vulnerability analysis of inverter control systems and the communication systems used to coordinate smart grids. Reference [9] presents an overview of existing attack detection approaches and outlines the benefits and limitations of knowledge-based and data-driven approaches. However, the methods in [9] only focus on general cyber-physical systems, and their performance in the context of IBRs requires further investigation. Reference [10] specifically outlines how photovoltaic systems are vulnerable to both sensor and communication-based attacks and classifies the detection approaches into model-based and data-driven techniques. Model-based detection validates measurements against predictions made by a physics-based model of the system and can detect attacks that inject false data. However, such methods are vulnerable to model-based attacks, e.g., stealthy attacks [6]. Data-driven detection methods train machine learning algorithms on historical data and can detect various attacks, but their performance generally lacks physical interpretation, which limits their application in safety-critical infrastructure, e.g., power systems [10].

Among the model-based methods, a dynamic watermarking approach [6], [7] shows potential for detecting false data injection attacks in IBRs. The watermarking approach superimposes a small noise (called a "watermark" signal) on the controla commands to detect cyber attacks in the IBR sensors. However, the watermark signal is injected solely for attack detection, and it may compromise the control performance of IBRs. This paper therefore proposes a novel approach that leverages renewable fluctuations to watermark the IBR measurements and identify measurements under cyber attacks. Compared with the conventional watermarking approaches from [6] and [7], the method developed in this paper does not require external noise injection and thus does not impact the IBR control performance.

The rest of this paper is organized as follows. Section II describes the dynamics of IBRs and points out the cyber vulnerability in the IBRs; Section III introduces the natural watermarking approach; Section IV tests the effectiveness of the proposed approach in a single IBR system; and Section V summarizes this paper and points out future work.

This work was supported by the U.S. National Science Foundation (NSF) under Award 2328205.

II. Problem Formulation

We use the solar-powered IBR shown in Fig. 1 to present a basic implementation of our natural watermarking, which can be extended to the IBRs powered by other types of renewable energy, e.g., wind. A variable DC voltage V_{DC} models the intermittent irradiation received by a solar panel. Due to hardware like DC-link capacitors and the slow change rate of irradiation, V_{DC} variations will always be well below the inverter switching frequency and there will be no significant interaction between V_{DC} and the switching dynamics. The average model of a two-level 3-phase voltage source inverter (VSI) is therefore used to simulate the terminal voltages [11]:

$$\mathbf{v}_{\mathrm{abc}} = \mathbf{m}_{\mathrm{abc}} \frac{V_{\mathrm{DC}}}{2} \tag{1}$$

where $\mathbf{m}_{\mathrm{abc}} \in \mathbb{R}^3$ is the voltage modulation index and V_{DC} is the DC-link voltage. The inverter output is then passed through an LCL filter and a closed loop control system is established through the measurements of the inverter current $\mathbf{i}_{\mathrm{f}}$, the capacitor voltage $\mathbf{v}_{\mathrm{c}}$, and the filter output current $\mathbf{i}_{\mathrm{o}}$. The feedback consists of a droop controller followed by a voltage and current controller. To reduce the necessary control complexity all 3-phase measurements are converted to the synchronous direct-quadrature frame (dq-frame) and the equations governing each component can be found in [12].

Without considering the switching dynamics of the inverter in Fig. 1, the dynamics of the inverter with the LCL filter can be described by

$$\dot{\mathbf{x}} = A\mathbf{x} + \frac{1}{2} B V_{\mathrm{DC}} \mathbf{m}_{\mathrm{abc}} \tag{2a}$$

$$\mathbf{y} = C\mathbf{x} \tag{2b}$$

where vector $\mathbf{x}$ collects the states of the three-phase output filter; matrices A, B, and C result from the filter dynamics; and $\mathbf{y} \in \mathbb{R}^9$ collects the three-phase currents $\mathbf{i}_{\mathrm{f}}$ and $\mathbf{i}_{\mathrm{o}}$, and the voltages $\mathbf{v}_{\mathrm{o}}$ of the filter. It is worth noting that the state-space model in (2a), (2b) contains a bilinear term $V_{\mathrm{DC}}\mathbf{m}_{\mathrm{abc}}$, and, therefore, it is nonlinear. We will exploit this observation later in Section III.

The attack symbol in Fig. 1 points out the location where an attacker can manipulate the actual measurements $\mathbf{y}$. By doing this, the attacker can control the inverter operation and cause the IBR to malfunction, harming both the inverter and the connected loads, and compromising grid security. In this article, we assume that the DC voltage measurements are secured, since the DC voltage sensors are less physically accessible and not directly used within the control loop, making them less likely to be a target for attackers.

III. Natural Watermarking Approach

In this section, we discuss in detail how the natural watermarking approach detects cyber attacks.

A. *Watermarking Measurements via Renewables to Detect Cyber Attacks*

Fig. 1 can be simplified by the feedback system shown in Fig. 2. The active watermarking approach injects an independent and identically distributed watermark signal $\mathbf{e}$, with known statistics onto the inverter control input $\mathbf{m}_{\mathrm{abc}}$ as shown in Fig. 2. This externally injected watermark then propagates through the system and appears in the sensor measurements $\mathbf{i}_{\mathrm{f}}, \mathbf{v}_{\mathrm{c}}$, and $\mathbf{i}_{\mathrm{o}}$. By checking the existence of the watermark signal in the sensor measurements via two statistical tests, a wide range of false data injection attacks in the measurements can be detected [13]. Note that detecting cyber attacks in the measurements requires a noise injection $\mathbf{e}$ which may compromise the control performance of the controller in Fig. 2.

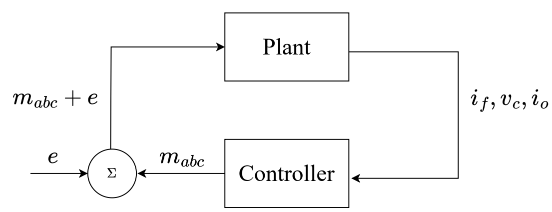


Fig. 2. Active watermarking approach.

The natural watermarking approach proposed in this paper aims to eliminate the need for injecting the noise $\mathbf{e}$ in the

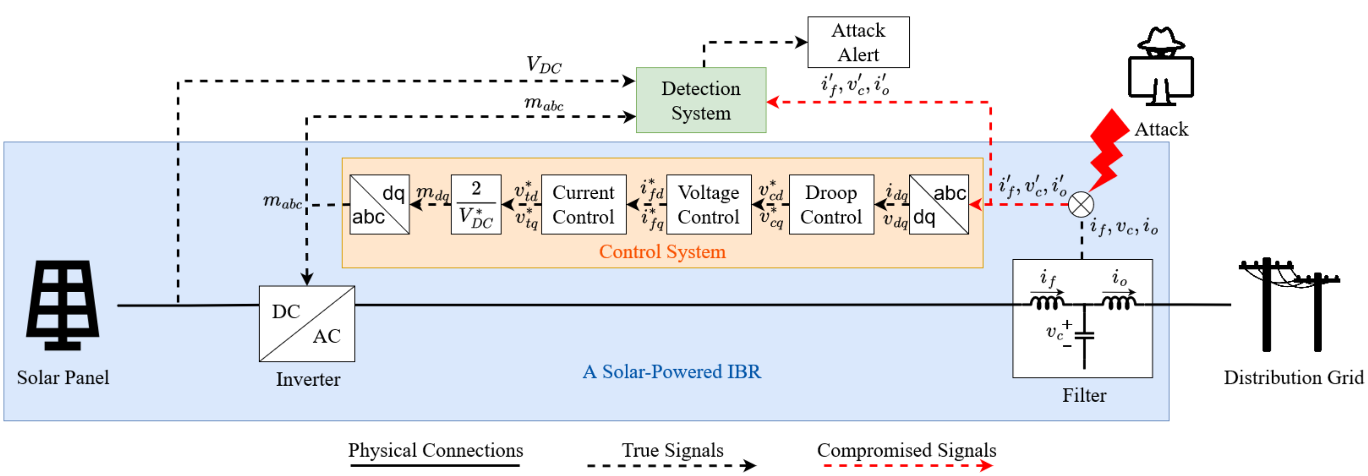


Fig. 1. System block diagram.

conventional watermarking approach. The proposed approach leverages the fluctuations of renewable energy generation, V_{DC}, as a *natural* watermark signal, as illustrated in Fig. 1. Changes in environmental conditions such as temperature and solar irradiation introduce a stochastic feature to variable V_{DC}. This makes it challenging for the attackers to reproduce this signal. The renewable fluctuation V_{DC} behaves like a natural watermark for the measurements and thus removes the need for externally injecting white noise $\mathbf{e}$.

The next question is how to check if the natural watermark, i.e., the renewable fluctuation V_{DC}, exhibits in the measurements $\mathbf{y}$. For the i-th sensor, this can be done by computing two types of indicators: moving average χ_{1i} and moving variance χ_{2i}. Denoted by $\mathbf{z}$ is the sensor measurements of $\mathbf{y}$. An FDI attack may cause $\mathbf{z} \neq \mathbf{y}$. Each element z_i in vector $\mathbf{z}$ corresponds to the measurement reported by sensor i which is one of measurements of the three-phase $\mathbf{i}_{\mathrm{f}}$, $\mathbf{v}_{\mathrm{c}}$, and $\mathbf{i}_{\mathrm{o}}$. We calculate the difference $\Delta z_i[k]$, between the corresponding elements of the measured and the predicted values of each sensor i:

$$\Delta z_i[k] = |z_i[k] - \hat{z}_i[k]| \tag{3}$$

where $\hat{z}_i[k]$ is the predicted measurement at k-th data point according to dynamics (2), V_{DC}, and $\mathbf{m}_{\mathrm{abc}}$. With Δz_i, the two indicators $\chi_{1i}[l]$ and $\chi_{2i}[l]$ at time l are computed in a sliding-window fashion. In Fig. 3, the red-dashed box shows the window at time l that collects W data points from time $l-W+1$ to time l. Using the data collected in the red-dashed box, we compute $\chi_{1i}[l]$ and $\chi_{2i}[l]$ by

$$\chi_{1i}[l] = \frac{1}{W} \sum_{k=l-W+1}^{l} \Delta z_i[k] \tag{4a}$$

$$\chi_{2i}[l] = \frac{1}{W} \sum_{k=l-W+1}^{l} (\Delta z_i[k] - \chi_{1i})^2. \tag{4b}$$

When $\eta \in \{1, 2, \ldots, W\}$ new data points are received, the data points in the red-dashed box are updated by removing η past data points and adding η new data points. The updated data points are collected by the blue-dashed window in Fig. 3. Then we recalculate $\chi_{1i}[l]$ and $\chi_{2i}[l]$ based on the updated data points in the the blue-dashed window. After repeating the above process, we will obtain a sequence of χ_{1i} and χ_{2i}. Once a cyber attack occurs at the i-th sensor, the values of χ_{1i} and χ_{2i} will start shooting up significantly compared to the no-attack scenario, making the attack detectable.

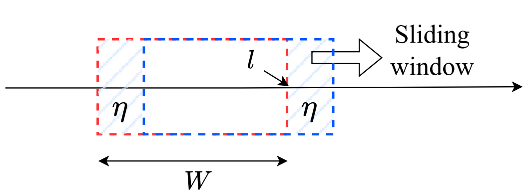
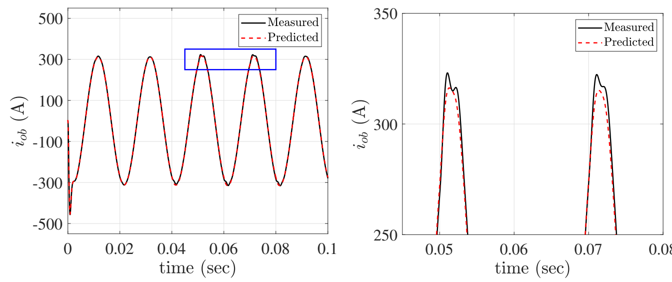


Fig. 3. Sliding window approach.

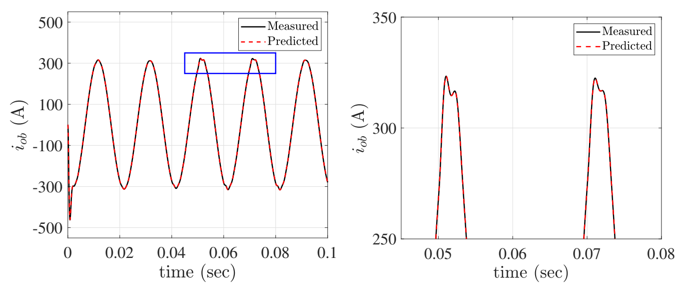
B. System Identification

The detection of cyber attacks using natural watermarking requires prior knowledge of the system inputs V_{DC}, and $\mathbf{m}_{\mathrm{abc}}$, and the plant model (2a) and (2b). This knowledge is necessary to obtain $\hat{z}_i[k]$ which is compared with the measured outputs $z_i[k]$ to detect cyber abnormalities. The plant model of the system can be obtained by analytically deriving the state-space model of the IBR system, as discussed in Section II. However, a third-party cybersecurity service provider may not know the IBR parameters to derive the state-space model. To overcome this challenge, we leverage the system identification technique in [14] to obtain the state-space model.



(a) (b)

Fig. 4. (a) Performance of system identification when using $\mathbf{m}_{\mathrm{abc}}$ and V_{DC} as the inputs and $\mathbf{i}_{\mathrm{o}}$ and $\mathbf{v}_{\mathrm{o}}$ as the outputs; (b) Zoomed-in plot of the area marked by the blue box.



(a) (b)

Fig. 5. (a) Performance of system identification when using $\mathbf{v}_{\mathrm{abc}}$ as the input and $\mathbf{i}_{\mathrm{o}}$ and $\mathbf{v}_{\mathrm{o}}$ as the outputs; (b) Zoomed-in plot of the area marked by the blue box.

One key question of system identification techniques is: what are the inputs and outputs? Without domain knowledge of IBR dynamics, a natural option is to consider $\mathbf{m}_{\mathrm{abc}}$ and V_{DC} as two separate inputs, and $\mathbf{i}_{\mathrm{o}}$ and $\mathbf{v}_{\mathrm{o}}$ as the outputs. To perturb the system we change V_{DC}, which has a mean of 800 V, by adding a normally distributed (Gaussian) fluctuation with a variance of 50 V at every 10 ms. Then we use the system identification toolbox in MATLAB to obtain the state-space model. Fig. 4 shows the system identification performance: while the predicted measurement can capture the general trend of the actual measurement, the predicted measurements do not capture the dynamics caused by the change of $\mathbf{m}_{\mathrm{abc}}$ and V_{DC}. Such a poor performance results from the non-linearity of dynamics (2).

However, if we consider $\mathbf{m}_{\text{abc}}V_{\text{DC}}$ as the inputs, system (2) becomes linear. With such an observation, we perform the system identification by choosing $\mathbf{m}_{\text{abc}}V_{\text{DC}}$ as inputs. The performance of the system identification is shown in Fig. 5. It can be observed that the predicted measurements capture not only the general trend of the measurements but also the transients due to the perturbation of $\mathbf{m}_{\text{abc}}$ and V_{DC}, resulting in a high prediction accuracy of over 99.9%.

IV. Case Study

In this section, the test system is briefly described and the natural watermarking approach is evaluated against noise injection attacks, replay attacks, and stealthy attacks.

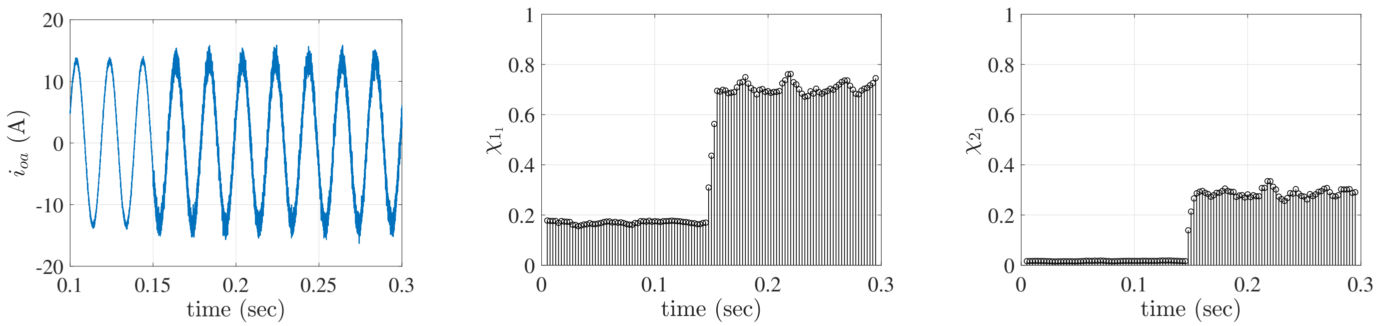
A. Test System Description

The output voltage of the PV system, V_{DC}, from Fig. 1 is modeled with a mean of 800 V and a normally distributed (Gaussian) fluctuation at every 10 ms. This fluctuation of V_{DC} creates a natural watermark that propagates to the system measurements and is essential for cyber attack detection. The solar PV system is then connected to an inverter modeled using the average model. The output of the inverter is connected to an LCL filter with the parameters $L_{\text{f}} = 1.35$ mH, $C_{\text{f}} = 50$ μF, $r_{\text{f}} = 0.1$ Ω, $L_{\text{c}} = 0.35$ mH, $r_{\text{c}} = 0.03$ Ω. A resistive load of 25 Ω is connected to the system.

B. Simulation Results

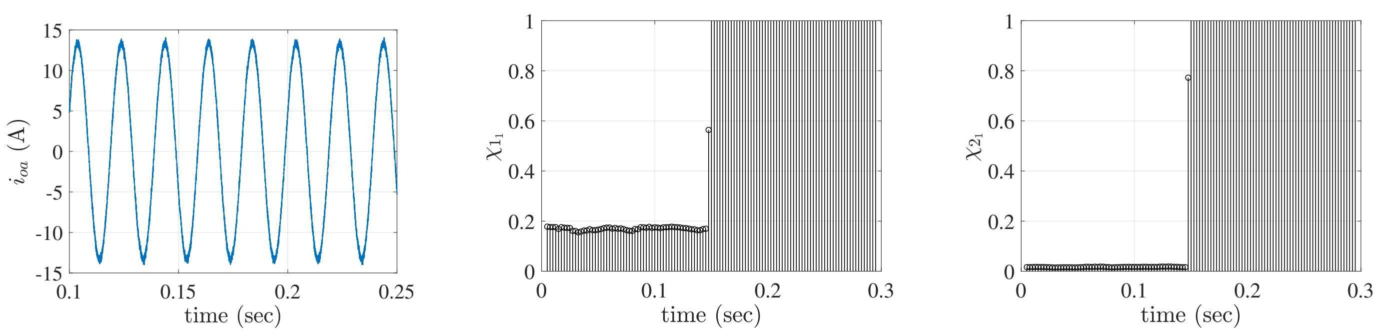
1) Noise Injection Attack: In this attack scenario, a Gaussian noise is added to the sensor measurements. This distorts the current and voltage measurements that are used in the control loop and affects the normal operation of the system. Fig. 6a shows how the filter output current in phase A is distorted under a noise injection attack starting at 0.15 s. The sudden increase in the indicators as seen in Fig. 6b and 6c clearly indicates the presence of the attack. This alerts the system operators about the system intrusion and the noise injection attack can be detected almost immediately.

2) Replay Attack: In this attack scenario, the attacker replaces the true system measurements with recorded measurements and keeps replaying the recording to evade detection. This is an advanced attack scenario where there is no visible distortion in the sensor measurements. Fig. 7a shows a replay attack launched on the filter output current in phase A at 0.15 s. This looks as if the system is healthy and is in normal operation conditions even if there is an attack present in the system. Although the current measurement does not indicate any distortions or abnormal behavior, the sudden increase in both χ_1 and χ_2 clearly indicates that there is an anomalous behavior in the system as seen in Fig. 7b and 7c. Thus, the replay attack can be detected in less than 1.5 cycles after the



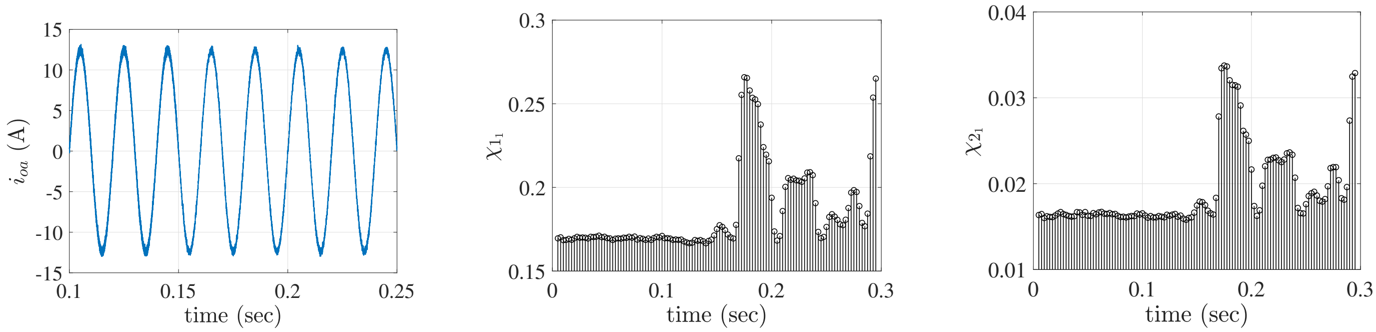
(a) (b) (c)

Fig. 6. Noise injection attack at 0.15 s and detection using the proposed attack indicators: (a) Waveform of i_{oa} before and after the attack; (b) Detection using χ_{1_1}, the attack indicator for sensor 1 which measures i_{oa}; (c) Detection using χ_{2_1}, the attack indicator for sensor 1 which measures i_{oa}.



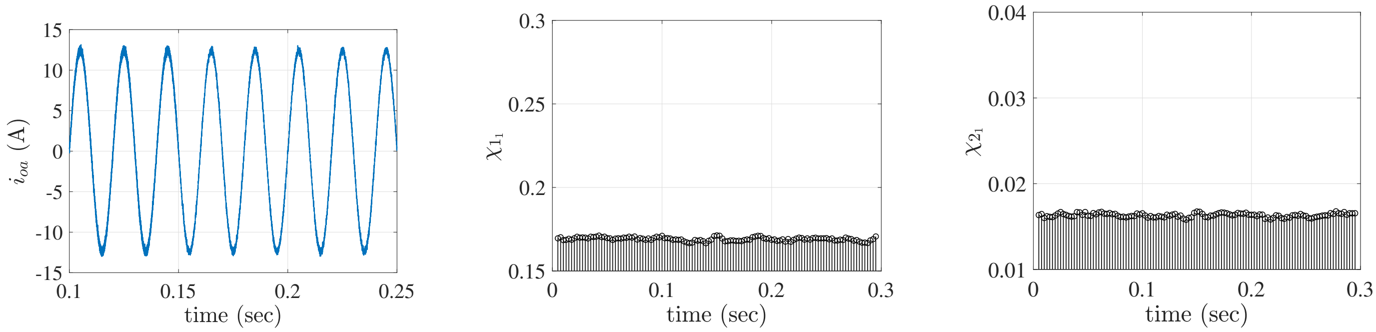
(a) (b) (c)

Fig. 7. Replay attack at 0.15 s and detection using the proposed attack indicators: (a) Waveform of i_{oa} before and after the attack; (b) Detection using χ_{1_1}, the attack indicator for sensor 1 which measures i_{oa}; (c) Detection using χ_{2_1}, the attack indicator for sensor 1 which measures i_{oa}.



(a) (b) (c)

Fig. 8. Detection of stealthy attack at 0.15 s due to the presence of the watermark signal in the sensor measurements: (a) Stealthy attack in i_{oa}; (b) Detection using χ_{1_1}, the attack indicator for sensor 1 which measures i_{oa}; (c) Detection using χ_{2_2}, the attack indicator for sensor 1 which measures i_{oa}.



(a) (b) (c)

Fig. 9. Undetected stealthy attack at 0.15 s due to the absence of the watermark signal in the sensor measurements: (a) Stealthy attack in i_{oa}; (b) Detection using χ_{1_1}, the attack indicator for sensor 1 which measures i_{oa}; (c) Detection using χ_{2_2}, the attack indicator for sensor 1 which measures i_{oa}.

attack using the proposed approach.

3) Stealthy Attack: In this attack scenario, the attacker predicts the behavior of the system based on the system model and measurements. With the predicted measurements $\hat{\mathbf{z}}$, the attacker forces the sensors to report $\hat{\mathbf{z}} + \mathbf{n}$ where $\mathbf{n}$ is a small noise different from the true noise. The response of the detection system under a stealthy attack in i_{oa} is shown in Fig. 8 and it can be seen that the attack is detected by the proposed attack indicators χ_1 and χ_2.

Next, we show why the natural watermark is necessary for detecting cyber attacks in our approach. When the natural watermark is not present in the sensor measurements which means that our watermark signal V_{DC} is a constant, the response of the detection system in such a scenario is shown in Fig. 9. It can be seen that the stealthy attack is not detected by the proposed attack indicators χ_1 and χ_2, as the indicators keep almost constant. Hence, this validates the necessity of the natural watermark signals in our algorithm for cyber attacks detection. Even if the solar PV output voltage, which is our watermark signal, remains constant for a while, our detection system would still be able to detect cyber attacks, once the PV output starts to fluctuate again.

V. Conclusion

This paper proposes a novel cyber attack detection method that leverages the natural fluctuations of renewable generation, e.g., PV panels, to watermark the IBR measurements and thereby detect cyber attacks in the system. Compared with conventional watermarking approaches, the proposed technique does not require an external noise injection as the watermark signal. Thus, it does not affect the IBR control performance. The attack detection is carried out using two types of indicators namely the moving average χ_1 and the moving variance χ_2. The performance of the proposed method is validated against noise injection attacks, replay attacks, and stealthy attacks. All these attacks are successfully detected by the proposed method in a short time. This approach can be extended to grid-connected multi-IBR systems powered by heterogeneous renewable resources. Future work will verify the results using hardware experiments under realistic renewable fluctuations.

References

[1] L. Xie, T. Huang, P. R. Kumar, A. A. Thatte, and S. K. Mitter, "On an information and control architecture for future electric energy systems," *Proceedings of the IEEE*, vol. 110, no. 12, pp. 1940–1962, 2022.

[2] K. E. Hemsley and D. R. E. Fisher, "History of industrial control system cyber incidents," 12 2018. [Online]. Available: https://www.osti.gov/biblio/1505628

[3] D. E. Whitehead, K. Owens, D. Gammel, and J. Smith, “Ukraine cyber-induced power outage: Analysis and practical mitigation strategies,” in *2017 70th Annual Conference for Protective Relay Engineers (CPRE)*, 2017, pp. 1–8.

[4] D. Kushner, “The real story of stuxnet,” *IEEE Spectrum*, vol. 50, no. 3, pp. 48–53, March 2013.

[5] T. Huang, D. Wu, and M. Ilić, “Cyber-resilient automatic generation control for systems of ac microgrids,” *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 886–898, 2024.

[6] H. A. J. Ibrahim and et al., “Detection of cyber attacks in grid-tied pv systems using dynamic watermarking,” *IEEE Transactions on Industry Applications*, vol. 60, no. 1, pp. 819–827, Jan.-Feb. 2024.

[7] W.-H. Ko, J. A. Ramos-Ruiz, T. Huang, J. Kim, H. Ibrahim, P. N. Enjeti, P. R. Kumar, and L. Xie, “Robust dynamic watermarking for cyber-physical security of inverter-based resources in power distribution systems,” *IEEE Transactions on Industrial Electronics*, vol. 71, no. 7, pp. 7106–7116, 2024.

[8] N. D. Tuyen, N. S. Quan, V. B. Linh, V. V. Tuyen, and G. Fujita, “A comprehensive review of cybersecurity in inverter-based smart power system amid the boom of renewable energy,” *IEEE Access*, vol. 10, pp. 35 846–35 875, 2022.

[9] S. Tan, J. M. Guerrero, P. Xie, R. Han, and J. C. Vasquez, “Brief survey on attack detection methods for cyber-physical systems,” *IEEE Systems Journal*, vol. 14, no. 4, pp. 5329–5339, Dec. 2020.

[10] J. Y. et al, “A review of cyber–physical security for photovoltaic systems,” *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 10, no. 4, pp. 4879–4901, Aug. 2022.

[11] A. Yazdani and R. Iravani, *Two-Level, Three-Phase Voltage-Sourced Converter*, 2010, pp. 115–126.

[12] N. Pogaku, M. Prodanovic, and T. C. Green, “Modeling, analysis and testing of autonomous operation of an inverter-based microgrid,” *IEEE Transactions on Power Electronics*, vol. 22, no. 2, pp. 613–625, 2007.

[13] B. Satchidanandan and P. R. Kumar, “Dynamic watermarking: Active defense of networked cyber–physical systems,” *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, 2017.

[14] L. Ljung, *System Identification: Theory for the User*, ser. Prentice Hall information and system sciences series. Prentice Hall PTR, 1999, [Online]. Available: https://books.google.com/books?id=nHFoQgAACAAJ.