Wei WANG, Zhenyong ZHANG, Xin WANG, Xuguo JIAO, 2025. Black-box adversarial attacks on deep reinforcement learning-based proportional–integral–derivative controllers for load frequency control. *Frontiers of Information Technology & Electronic Engineering*, 26(11):2128-2142. https://doi.org/10.1631/FITEE.2401021

Black-box adversarial attacks on deep reinforcement learning-based proportional–integral–derivative controllers for load frequency control

Key words: Adaptive controller; Deep reinforcement learning; Load frequency control; Adversarial attacks

Corresponding author: Zhenyong ZHANG
E-mail: zhangzy@gzu.edu.cn
ORCID: https://orcid.org/0000-0003-0950-1525

Motivation

Although deep reinforcement learning offers superior adaptability for load frequency control compared to traditional proportional–integral–derivative (PID), its deployment is hindered by inherent vulnerabilities to adversarial attacks, which are predominantly studied under unrealistic “white-box” assumptions, where attackers possess full model knowledge. Motivated by the critical gap in analyzing realistic “black-box” threats, where model parameters are unknown, and the challenge that enhancing robustness often degrades standard control performance, this study aims to expose these vulnerabilities using zeroth-order optimization (ZOO) for gradient estimation and develop a defense strategy that balances security with stable operational efficiency.

Main idea

- A detailed analysis of the vulnerability of the deep reinforcement learning (DRL)-based adaptive controller is conducted from the attacker's perspective.
- A black-box attack is designed by incorporating the ZOO method in adversarial attacks, enabling effective attacks on DRL-based adaptive controllers.
- A new adversarial training paradigm is proposed that enhances the robustness of the DRL-based adaptive controller while maintaining control performance.

Method

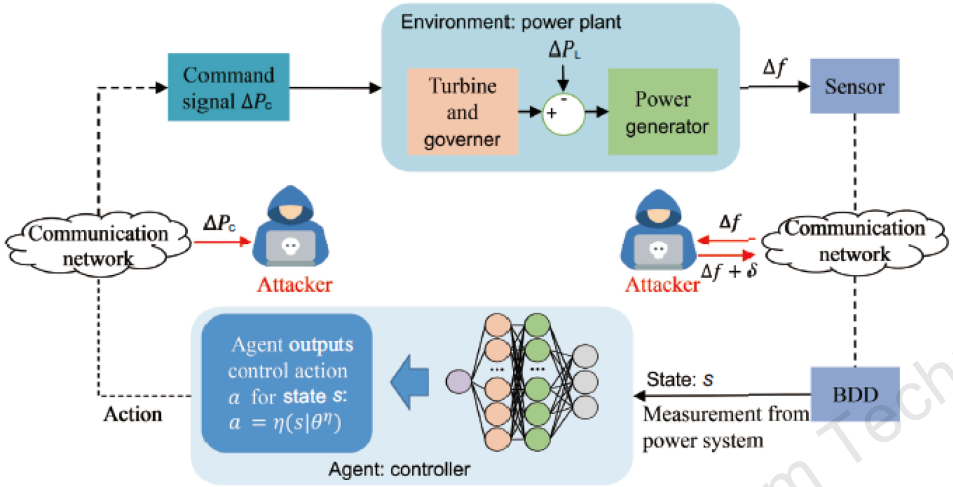


Fig. 2 DRL-based LFC for single-area system under the adversarial attacks. BDD: bad data detection

$$\begin{cases} x_{n+1}^{\text{adv}} = \Pi_{x_0+\delta}(x_n^{\text{adv}} - \alpha \text{sign}(\nabla_{x_n^{\text{adv}}} Q(s_0, \eta(s_n^{\text{adv}})))), \\ x_n^{\text{adv}} = x_0, \text{ if } n = 0, \end{cases}$$

$$\nabla_x Q(s, \eta(s)) \approx \mathbb{E}\left[\frac{Q(s, \eta(s + \epsilon u)) - Q(s, \eta(s))}{\epsilon} u\right]$$

- Attack Scenario
  - Man-in-the-Middle: The attacker intercepts communication traffic between the sensors and the control center.
  - Target: Manipulates the frequency deviation signal Δf by injecting adversarial perturbations δ.
- Core Challenge & Solution
  - Challenge: The DRL controller operates as a black-box with unknown parameters, making direct gradient computation impossible.
  - Solution: Uses the ZOO method to estimate gradients by querying inputs and outputs without requiring model knowledge.

Results

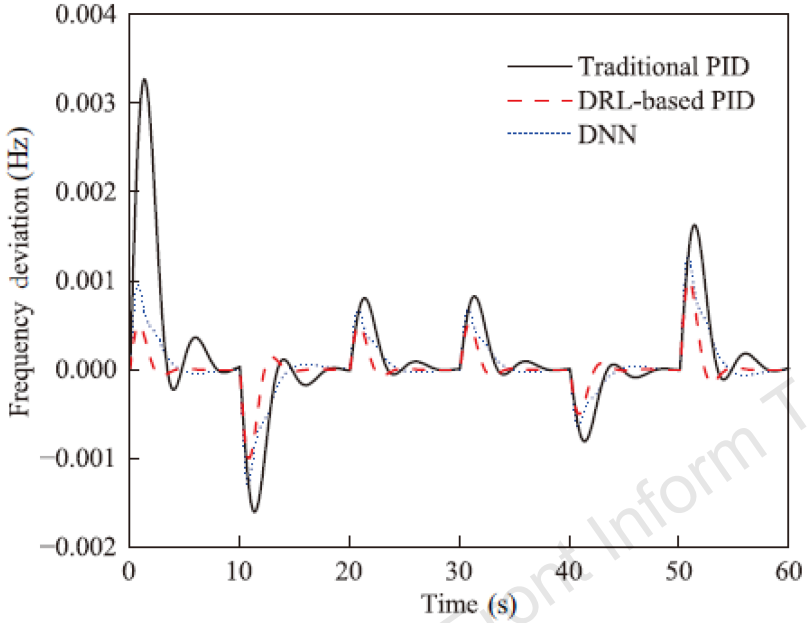


Fig. 4 Dynamic frequency deviation under step disturbance conditions

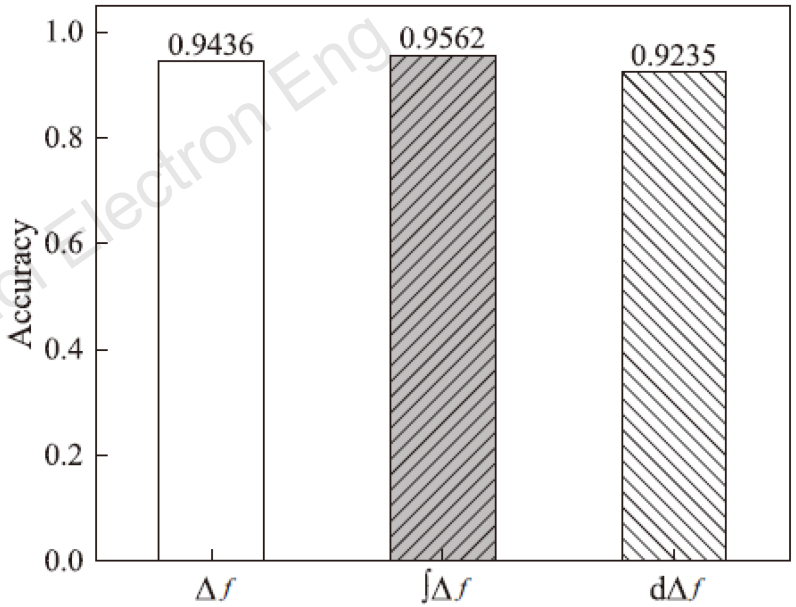


Fig. 5 Accuracy of the ZOO method in computing the gradient signs of the three state variables

Results

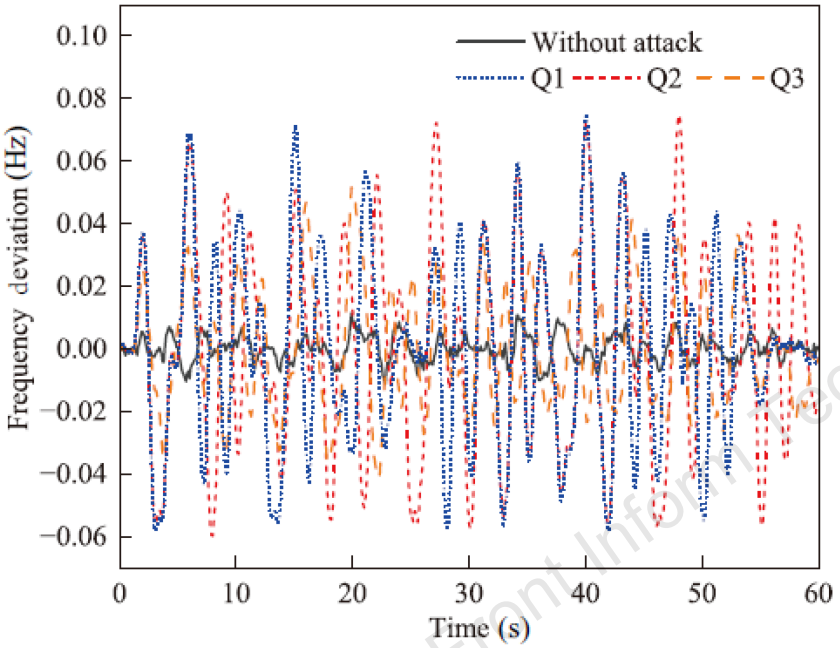


Fig. 6 Impact of different critic Q-network structures on adversarial attacks

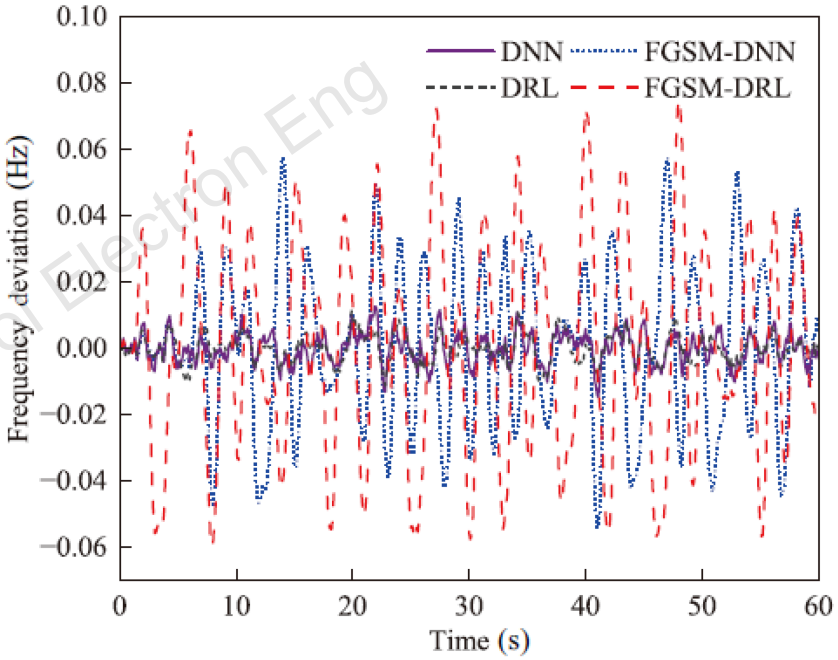
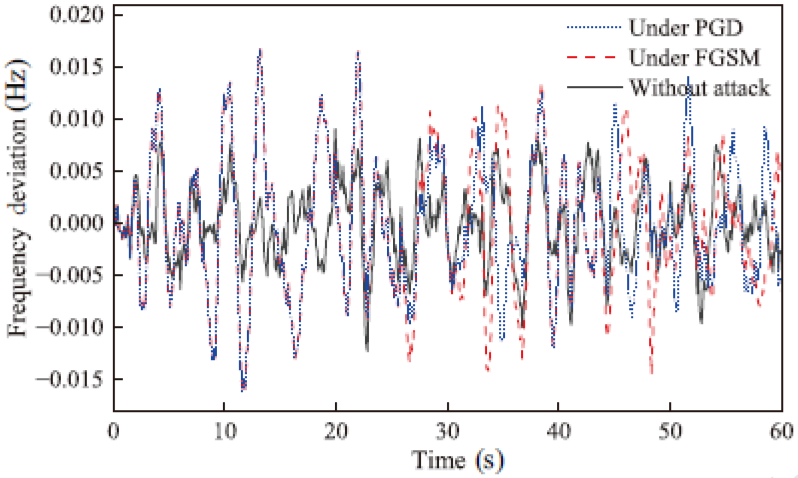
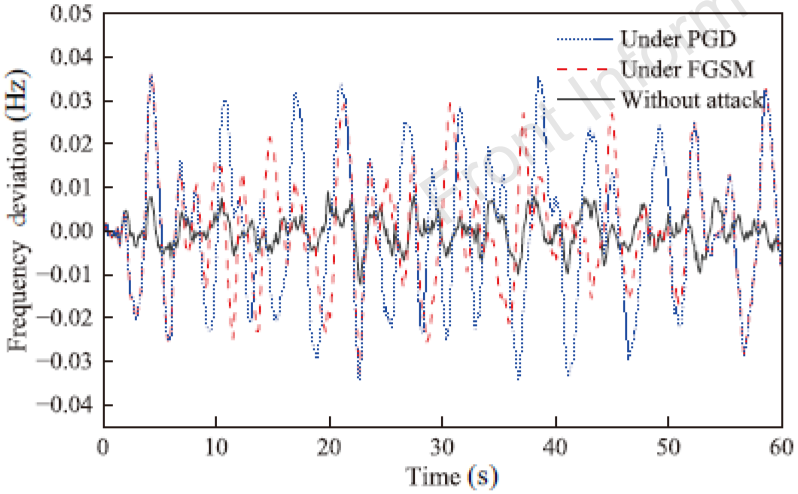


Fig. 7 The control performance of the DNN and DRL controllers under FGSM attacks

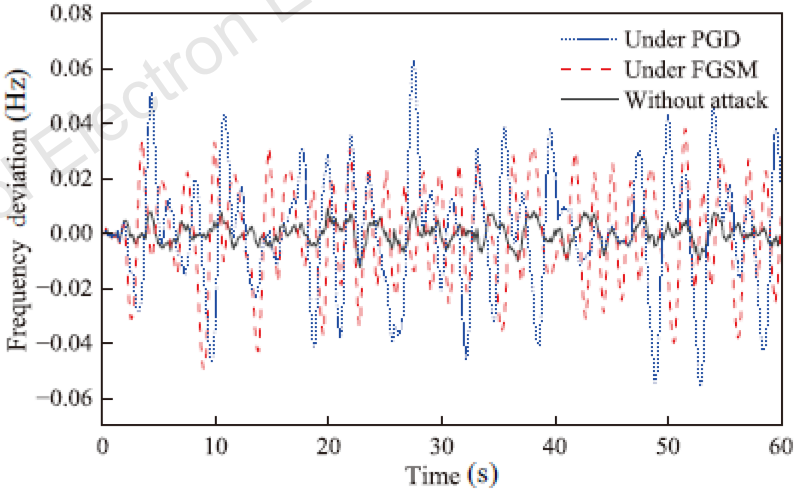
Results



(a)



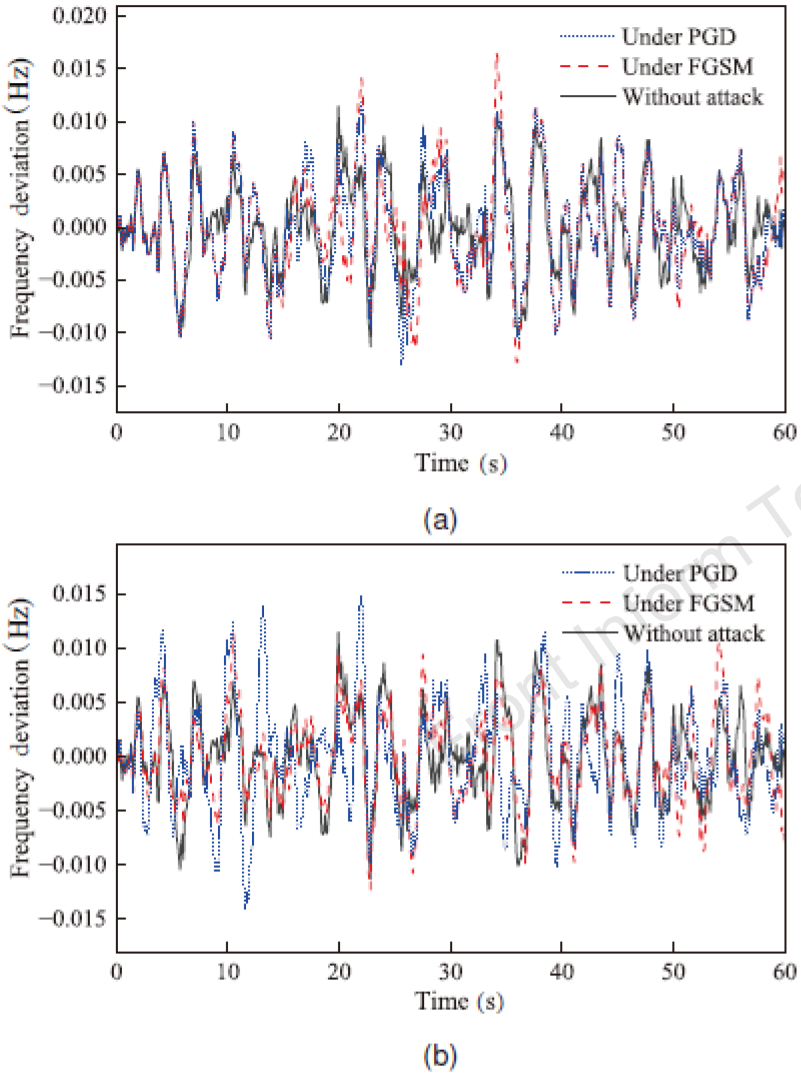
(b)



(c)

Fig. 8 Comparing the performance of the DRL-based controller under different attack types and varying perturbations: (a) $\sigma = 0.03$; (b) $\sigma = 0.06$; (c) $\sigma = 0.12$

Results



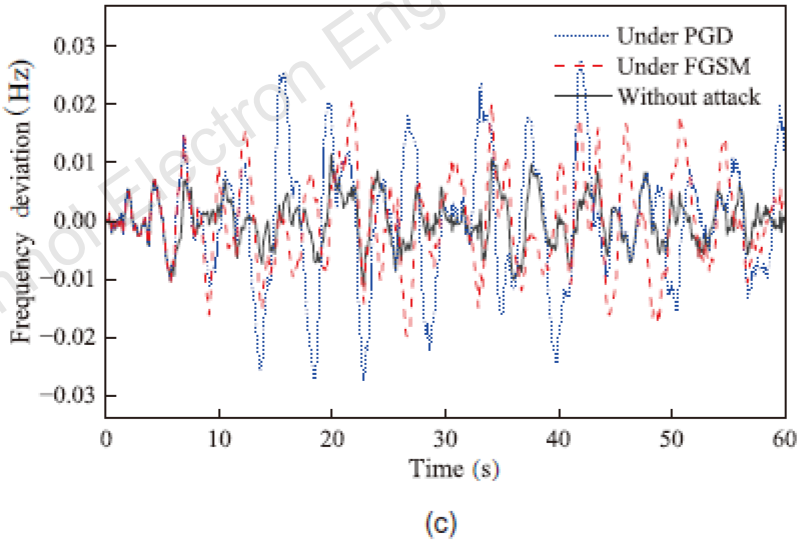


Fig. 9 The performance of the DRL-based controller after adversarial training under adversarial attacks: (a) $\sigma = 0.03$; (b) $\sigma = 0.06$; (c) $\sigma = 0.12$

Results

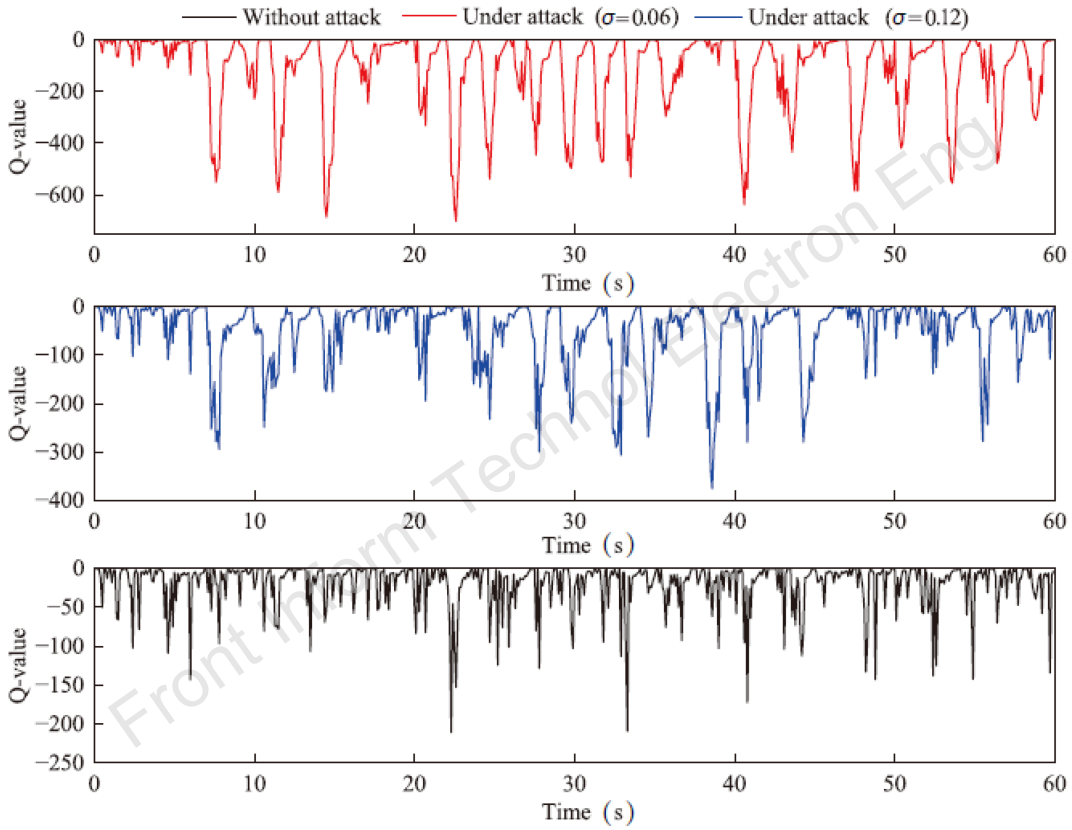


Fig. 10 The variation of Q-values under different attack intensities and no-attack. References to color refer to the online version of this figure

Results

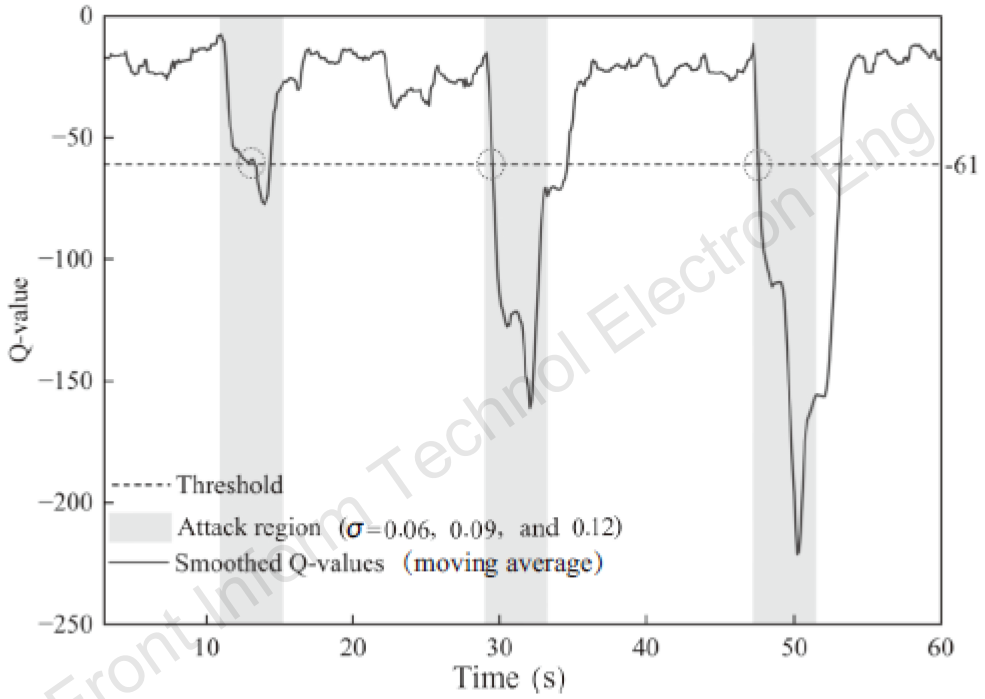
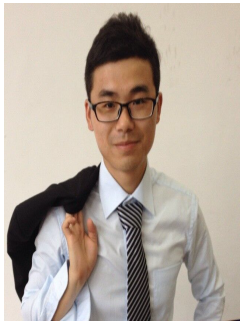


Fig. 11 Attack detection based on smoothed Q-value fluctuations under different attack intensities

Conclusions

This paper highlights the vulnerability of deep reinforcement learning-based load frequency control systems to black-box adversarial attacks, demonstrating that attackers can effectively compromise system stability using the zeroth-order optimization method without prior knowledge of model parameters. To mitigate this threat, the study proposes a hybrid adversarial training strategy that incorporates both normal and adversarial samples, alongside a real-time detection mechanism based on *Q*-value fluctuations. Simulation results confirm that these defense measures significantly enhance the robustness and security of the controller against malicious cyberattacks while maintaining effective performance under normal operating conditions.

Wei WANG received the bachelor's degree from Shanxi University of Finance and Economics, Taiyuan, China, in 2019. He is currently pursuing a master's degree at the College of Computer Science and Technology, Guizhou University, Guiyang, China. His research interests include power system security and machine learning.



Zhenyong ZHANG received the bachelor's degree from Central South University, Changsha, China, in 2015, and the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2020. He is currently a professor with the College of Computer Science and Technology, Guizhou University, Guiyang, China. His research interests include cyber-physical system security, applied cryptography, and machine learning security.

Xin WANG received the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou. China, in 2020. From 2018 to 2019, he was a visiting Ph.D. student at the Department of Computer Science, Tokyo Institute of Technology, Yokohama, Japan. He is currently with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, and also with Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science, Jinan. His research interests include computation and optimization, data privacy, and cyber-physical systems.

Xuguo JIAO received the B.S.degree in automation from Shandong Agricultural University, Tai'an, China, in 2012, the M.S.degree in control science and engineering from Chongqing University, Chongqing, China, in 2015, and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2020. He is currently an associate professor with the School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China. His research interests include wind power systems and intelligent control.