IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Deep Reinforcement Learning-Based Control Strategy with Direct Gate Control for Buck Converters

N. Katayama[1], Member, IEEE
[1]Tokyo University of Science, Chiba 2788510 Japan

Corresponding author: Noboru Katayama (e-mail: katayama@rs.tus.ac.jp).

ABSTRACT This paper proposes a deep reinforcement learning (DRL)-based approach for directly controlling the gate signals of switching devices to achieve voltage regulation in a buck converter. Unlike conventional control methods, the proposed method directly generates gate signals using a neural network trained through DRL, with the objective of achieving high control speed and flexibility while maintaining stability. Simulation results demonstrate that the proposed direct gate control (DGC) method achieves a faster transient response and stable output voltage regulation, outperforming traditional PWM-based control schemes. The DGC method also exhibits strong robustness against parameter variations and sensor noise, indicating its suitability for practical power electronics applications. The effectiveness of the proposed approach is validated via simulation.

INDEX TERMS DC-DC converter, optimal control, direct gate control, deep reinforcement learning, artificial intelligence

I. INTRODUCTION

Recently, various approaches have been proposed to apply deep reinforcement learning (DRL) techniques to the control of power electronic converters. DRL has been applied in various fields, where it learns optimal control strategies through trial-and-error interactions with the environment. DRL is capable of handling complex control targets, and by designing the reward function as the objective, it can achieve multiple goals simultaneously. Power electronics circuits are also becoming increasingly complex, requiring not only dynamic response but also high energy efficiency, reduced EMI, and other objectives to be satisfied at the same time. Under such conditions, designing controllers based on conventional classical control methods becomes challenging. By employing DRL, it is possible to train an agent through extensive trial-and-error in a simulation environment, thereby eliminating the need for detailed manual design expertise of complex controllers.

Hajihosseini et al. [1] utilized RL to optimize the parameters of a PI controller in a buck-boost converter, demonstrating superior tracking performance to the reference signal compared to terminal sliding mode control. Previous studies have investigated the modification of the duty ratio of gate signals in switching devices using DRL. For example, Gheisarnejad et al. [2] proposed a method to enhance the control performance of power electronic converters by augmenting the control signal of a conventional PID controller with a compensation signal obtained through deep deterministic policy gradient. Their approach demonstrated improved response speed and stability, as well as enhanced robustness against dynamic variations in the operating conditions. Cui et al. [3] proposed a model-free DRL-based control method to address voltage instability in DC-DC buck converters caused by fluctuations in constant power loads. In their study, the DRL model outputs discrete values representing the change in the duty ratio. Other applications of DRL in the field of power electronics have been reported not only for simple buck converters but also for dual active bridge converter [4], motor controller [5] , inverters [6] and multiport DC-DC converters [7], [8].

Several studies have investigated the control of the duty ratio in power electronic converters using DRL. Lee et al. [9] employed a model trained via the Soft Actor-Critic (SAC) algorithm to directly regulate the PWM duty ratio of a DC–DC buck converter, demonstrating faster transient response compared to conventional methods such as PI control and model predictive control (MPC). Rajamallaiah et al. [10] also applied DRL to control the duty ratio of a buck converter and evaluated multiple DRL algorithms, concluding that Twin Delayed Deep Deterministic Policy Gradient (TD3) outperformed others in terms of control

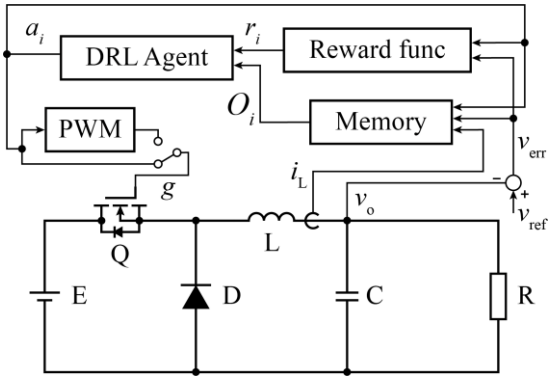

FIGURE 1. Process of anomaly detection using an autoencoder.

TABLE 1 Circuit parameters of the buck converter.

Parameter	Value
Input voltage, E	20 V
Inductance, L	47 μH
Inductor ESR, R_L	10 mΩ
Capacitance, C	470 μH
Capacitor ESR, R_C	100 mΩ
Load resistance, R	1-∞ Ω

performance. In addition, DRL-based duty radio control strategies for DC–DC converters have also been explored in [7], [11], [12].

While most prior work focuses on duty ratio modulation per PWM cycle, DRL has shown promise in more complex control problems beyond power electronics, such as robotics and video game environments. Among these studies, only Qashqai et al. [13] directly learned and controlled the on/off states of individual switching devices, rather than computing a duty ratio for PWM generation. Their DRL-based approach, applied to a three-level neutral point clamped (NPC) converter, exhibited robustness against parameter variations in passive components such as inductors and capacitors, and maintained stable operation even under measurement noise.

Building on these insights, it is anticipated that DRL can be employed to directly control the on/off states of switching devices in converter topologies such as buck converters, without relying on PWM generation. This enables more flexible and faster control by eliminating the constraints imposed by a fixed PWM period. Specifically, a control strategy can be envisioned in which circuit voltages and currents are sampled at a higher rate than the switching frequency, and the appropriate switching actions are determined by the DRL agent at each sampling instant without the conventional duty ratio calculation. Since the control target in Qashqai et al.'s study was a DC/AC conversion circuit, the control interval was 20 μs; however, in buck converters, a much higher control frequency is required. Furthermore, a comparative evaluation with DRL-based control methods that output the duty ratio in previous studies is also required.

TABLE 2 Hyperparameters for DRL.

Parameter	Value
Discount factor	0.999 / 0.99 (for PWM)
Clip range	0.2
Learning rate	3×10^{-4}
Number of epochs	10
Batch size	64
Rollout steps	2048
Train steps	1,000,000
Policy network	[64, 64]
Critic network	[64, 64]
Reward parameters, $\alpha, \beta, \zeta, \delta$	0.2, 0.004, 0.1, 4

In this study, a DRL-based control method—hereafter referred to as direct gate control (DGC)—is proposed to directly generate gate signals for a buck converter, allowing the agent to determine the switching state of the MOSFET at each control step without relying on PWM generation. The proposed DGC method is evaluated through simulation studies, demonstrating its effectiveness in achieving voltage regulation while simplifying the overall control architecture.

II. METHODOLOGY

A. SYSTEM DESCRIPTION

The target system in this study is a conventional buck converter, consisting of a single MOSFET as the main switching device, a freewheeling diode, an inductor, an output capacitor, and a resistive load as shown in Figure 1. This topology is widely used in DC–DC conversion due to its simplicity. In this study, the buck converter is selected not only for its simplicity but also to focus on the control aspects of DRL. The primary objective is to regulate the output voltage to a desired reference level by appropriately controlling the switching state of the MOSFET. The inductance and the capacitance of the circuit include the equivalent series resistances to approximate the behavior of the actual circuit.

To implement and evaluate the proposed control strategy, the converter is modeled and simulated in PLECS, a simulation environment specifically designed for power electronic systems. All components are treated as ideal except for the inclusion of equivalent series resistance (ESR) in the inductor and capacitor. At each control step, PLECS exports the current system observation—such as output voltage and inductor current—to the agent for computing the next gate signal.

B. CONTROL FRAMEWORK

The proposed framework eliminates the PWM generation stage and allows the DRL agent to directly determine the gate signal of the switching device in a binary signal. By doing so,

the control policy gains more flexibility in responding to system dynamics without being constrained by a predefined modulation scheme. Moreover, this architecture simplifies the overall control structure and opens the possibility for higher-speed and more adaptive switching behavior.

The control period is set to 1 μs, which defines both the state observation interval and the gate signal update rate. At each time step, the current system state—such as output voltage and inductor current—is sampled and passed to the DRL agent, which returns a discrete action representing the next gate state.

C. DEEP REINFORCEMENT LEARING ALGORITHM

The control policy for the proposed strategy is trained using the Proximal Policy Optimization (PPO) algorithm [14], a model-free, on-policy reinforcement learning method. PPO has gained popularity due to its ability to achieve stable and sample-efficient learning while being relatively easy to implement and tune. It adopts an actor–critic architecture, in which the policy and the value function are represented by separate neural networks and optimized simultaneously. Another reason for selecting the PPO algorithm is its ability to handle both continuous and discrete action spaces. This enables a fair comparison between the proposed DGC method and the conventional PWM-based approach using the exact same algorithmic framework.

In this study, the action space is defined as a binary set representing the on/off states of the switching device. During training, actions are sampled stochastically from the policy distribution. During evaluation, a deterministic policy is used by selecting the action with the highest probability (i.e., the most probable action). This ensures reproducible behavior when comparing performance with baseline methods.

The neural networks for both the policy and value functions consist of two fully connected layers, each followed by a ReLU activation function. The output layer of the policy network produces the logits for a categorical distribution over the two discrete actions. The hyperparameters used for PPO training are summarized in Table 2.

D. DEFINITION OF OBSERVATION, ACTION AND REWARD SPACES

The control problem is formulated as a discrete-time Markov decision process, in which the agent receives the current system observation, selects a control action, and obtains a scalar reward at each control interval.

The observation vector at each time step consists of the recent history of the voltage error between the converter output and the reference voltage, the instantaneous inductor current, and the switching device's on/off states as follows:

$$O_t = \begin{Bmatrix} v_{err,t}, v_{err,t-1}, v_{err,t-2}, \cdots, v_{err,t-N-1}, \\ i_{L,t-1}, i_{L,t-2}, \cdots, i_{L,t-N-1}, a_{t-1}, a_{t-2}, \cdots, a_{t-N-1} \end{Bmatrix} \quad (1)$$

where $v_{err,t}$, $i_{L,t}$ and a_t denotes the output voltage error, inductor current and switching state at time step t, respectively.

Each value is recorded over the previous ten sampling steps. This temporal history is incorporated to provide the agent with short-term dynamic information relevant to recent switching behavior and state transitions, thereby enhancing the observability of internal converter dynamics.

The action space is defined as a discrete set with two possible values, on and off states of the switching device. At every control step, the agent selects one of the two actions. To account for the processing delay introduced by the ADC and the policy network inference, the selected action is applied to the switching device with a one-step delay. In the PWM control strategy, the action space is defined as a one-dimensional continuous variable, representing the duty ratio within the range of 0.0 to 1.0.

The reward function is defined to promote accurate output voltage regulation and penalize excessive switching. A general form of the reward used during training is given by:

$$r_{t,1} = \frac{\alpha}{|v_{err,t}|+\epsilon} - \zeta|v_{err,t}| - \beta \quad (2)$$

$$r_{t,2} = -\delta|a_t - a_{t-1}| \quad (3)$$

$$r_t = r_{t,1} + r_{t,2} \quad (4)$$

The other constants are listed in Table 2. The behavior of circuit controlled by DRL can be flexibly designed through the reward function.

E. TRAINING ENVIRONMENT

Each training episode consists of 2,000 time steps, corresponding to a simulation duration of 2 ms with a control interval of 1 μs. To promote generalization and ensure the agent experiences a wide range of operating conditions, the initial state for each episode is randomized. Specifically, the initial values of the inductor current, output capacitor voltage, and load resistance are independently sampled from predefined ranges. The load resistance has a constant value during each episode. This setup allows the agent to encounter diverse transient and steady-state behaviors during training.

F. BASELINE FOR COMPARISON

To evaluate the effectiveness of the DGC control, a baseline method is implemented in which the PWM duty ratio is directly controlled by a DRL agent. The same PPO algorithm and neural network architecture and the state vector are used for the baseline to ensure a fair comparison. In the baseline configuration, the agent outputs a continuous-valued duty ratio at each control step. This value is then applied to a conventional fixed-frequency PWM generator to produce the corresponding gate signal. The PWM period and control interval are set to 10 μs. Since the control interval is longer than the DGC method by ten times, the reward value is multiplied by 10 to account for the difference in control frequency. The discount rate is also adjusted to 0.99 to maintain consistency with the DGC method.

III. RESULTS AND DISCUSSION

A. DYNAMIC RESPONSE TO LOAD STEP CHANGES

The fundamental behavior of the proposed DRL-based DGC was first verified under nominal conditions. Figure 2 compares the response of the proposed method (blue line) and the conventional PWM control (orange line) under a load step disturbance from 15 Ω to 1 Ω at 0.0 ms. The top two graphs depict the gate signal and duty ratio, the middle plots show the inductor current, and the bottom plots display the output voltage (solid line) with the reference voltage (dotted red line). Since the concept of a duty ratio does not directly apply to DGC, an effective duty ratio, calculated by applying a digital low-pass filter to the gate signal, is shown for illustrative purposes.

After the output voltage initially drops to approximately 13.8 V due to rapid load current increase, the DGC method quickly adapts by adjusting the gate signal. The output voltage starts to rise at 0.06 ms, reaching the reference voltage within 0.15 ms. The gate signal stays high for 0.14 ms after the load change, subsequently, the gate signal alternates between on and off, bringing the output voltage asymptotically toward the reference voltage. Notably, the gate signal transitions are not constrained by a fixed switching frequency, providing greater control flexibility.

However, the PWM control method exhibits a slower response. The output voltage drops to 13.2 V and recovers to the reference voltage after 0.2 ms. The output voltage overshoots to 15.2 V and oscillates before settling around 15 V. The duty ratio is fixed at 1.0 for 0.13 ms, and adjusted to settle the output voltage to the reference voltage. Although the response is adequate, the performance shows poorer compared with DGC due to control delay and the fixed switching frequency.

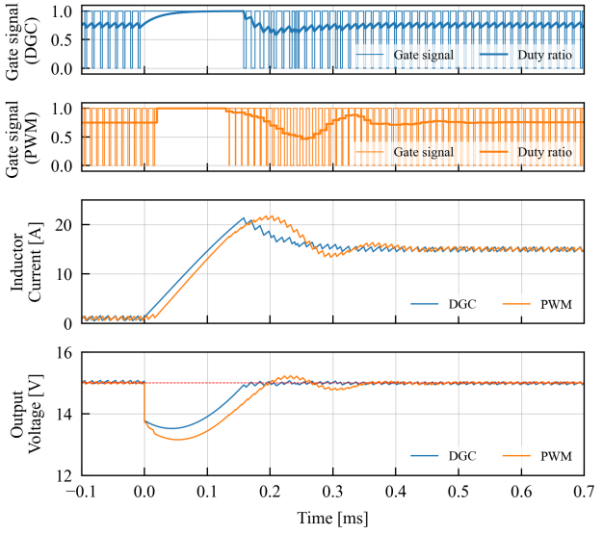

FIGURE 2. Comparison between the proposed DGC method and conventional PWM control in response to a load step change under nominal circuit parameters.

B. ROBUSTNESS EVALUATION UNDER PARAMETER VARIATIONS

To evaluate the robustness of DGC against parameter variations due to its flexible control scheme, dynamic response was evaluated varying important circuit parameters such as inductance and equivalent series resistance (ESR) of the

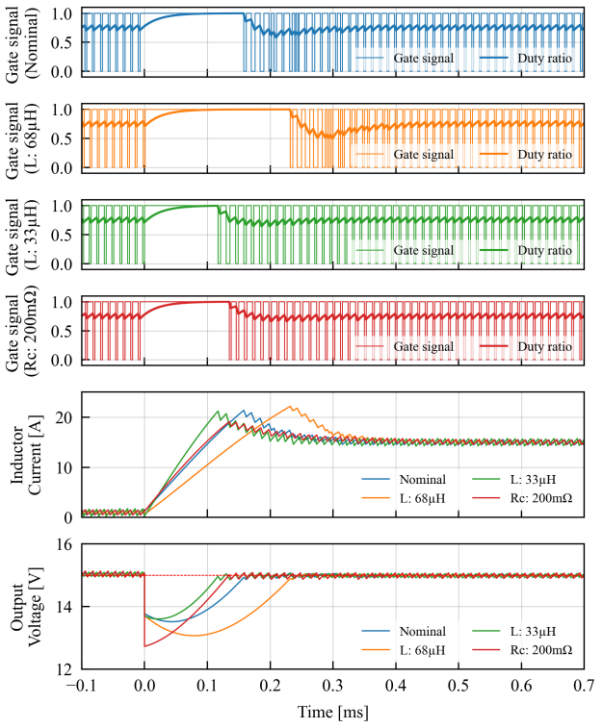

FIGURE 3. Response under variations in inductance and ESR of the capacitor.

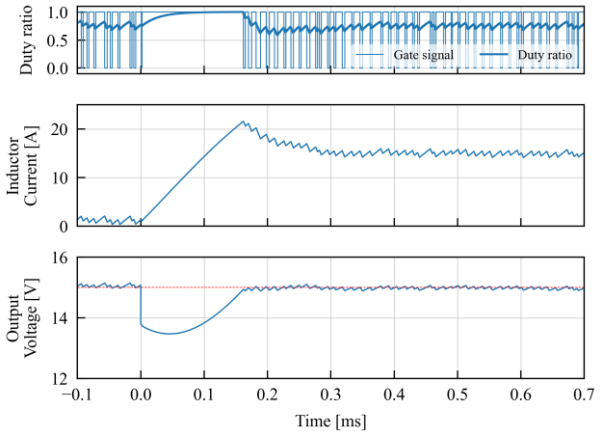

FIGURE 4. Response under nominal circuit parameters with injected measurement noise.

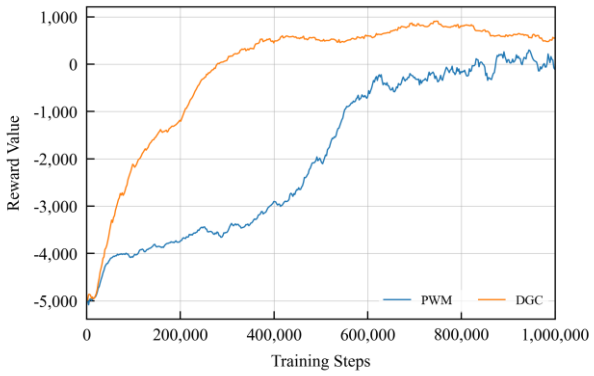

FIGURE 5. Comparison of reward evolution during training for DGC and PWM methods.

output capacitor. The inductor value is typically affected by the manufacturing process, while the ESR of the output capacitor can change due to aging or temperature effects.

Figure 3 shows the output voltage response when the inductor value is changed to 33 µH and 68 µH, and the ESR of the output capacitor is doubled from 100 mΩ to 200 mΩ, respectively. For all cases, the output voltage recovered to the reference voltage in a stable manner. Although the recovery time is slightly changed for replaced circuit parameters, all the cases show that the output voltage remains well-regulated.

C. ROBUSTNESS AGAINST SENSOR NOISE

Sensor noise was introduced to the output voltage measurement to evaluate further robustness. The noise was modeled as a Gaussian white noise with a standard deviation of 0.01 V for the output voltage and 0.1 A for the inductor current. Figure 4 shows the output voltage response under this noisy condition. The DGC method successfully maintained stable output voltage regulation despite the presence of noise, demonstrating robustness against measurement uncertainties. The switching behavior responds to the noisy measurements, showing the agent's ability to adapt quickly to perceived changes in the output voltage while maintaining overall stability.

D. TRAINING PERFORMANCE ANALYSIS

Figure 5 illustrates the learning curves of the proposed DGC and the PWM method in terms of reward values over training steps. Both methods started from a significantly negative reward due to initial random exploration. However, the DGC approach exhibited a much faster convergence speed, reaching a stable reward values at approximately 400,000 training steps. In contrast, the PWM-based control demonstrated slower learning progress. The reward value remained below –3,000 until about 400,000 training steps, and stable reward values were achieved after approximately 600,000 steps.

It should be noted, however, that the definition of a training step differs between the two approaches. In the DGC case, one step corresponds to a switching period of 1 µs, while in the PWM-based method, one step corresponds to 10 µs. Consequently, although DGC achieves convergence in fewer steps, the actual wall-clock time required for convergence is shorter in the PWM case.

The faster convergence of DGC in terms of training steps can be attributed to its direct and discrete action representation, which reduces the complexity of the state–action mapping compared to the continuous duty-ratio representation of PWM. In DGC, the agent directly determines the switching states, allowing the reward function to provide clearer feedback and accelerating policy optimization.

These results indicate that while DGC improves learning efficiency per step due to its direct control formulation, practical implementation requires consideration of the increased computational demand associated with its higher interaction frequency.

IV. CONCLUSIONS

This study proposed a novel DGC method for buck converters using deep reinforcement learning. The DGC method directly generates gate signals for the switching device, eliminating PWM generation. Simulation results demonstrated that the DGC method achieves fast transient response and stable output voltage regulation, outperforming traditional PWM-based control methods. Additionally, the DGC method exhibited robustness against parameter variations and sensor noise, confirming its potential for practical applications in power electronics.

Beyond these contributions, several challenges remain for future research. First, experimental validation on hardware prototypes is essential to confirm the feasibility of DGC under real-world operating conditions, including switching losses, device non-idealities, and thermal constraints. Second, while the present study considered a single converter topology, extending the DGC framework to multi-phase converters, bidirectional converters, and other topologies could further broaden its applicability. Third, the computational burden of real-time DRL inference must be carefully addressed through lightweight neural network architectures or hardware acceleration (e.g., FPGA or DSP implementations). Finally, ensuring long-term stability, safety constraints, and fault tolerance in the DGC framework remains an open issue that will be critical for industrial adoption.

These directions highlight the promising potential of DGC as a next-generation control paradigm in power electronics, while underscoring the need for continued investigation toward practical deployment.

REFERENCES

[1] M. Hajihosseini, M. Andalibi, M. Gheisarnejad, H. Farsizadeh, and M.-H. Khooban, "DC/DC Power Converter Control-Based Deep Machine Learning Techniques: Real-Time Implementation," *IEEE Trans. Power Electron.*, vol. 35, no. 10, pp. 9971–9977, Oct. 2020, doi: 10.1109/TPEL.2020.2977765.

[2] M. Gheisarnejad, H. Farsizadeh, and M. H. Khooban, "A Novel Nonlinear Deep Reinforcement Learning Controller for DC–DC Power Buck Converters," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 6849–6858, Aug. 2021, doi: 10.1109/TIE.2020.3005071.

[3] C. Cui, N. Yan, B. Huangfu, T. Yang, and C. Zhang, "Voltage Regulation of DC-DC Buck Converters Feeding CPLs via Deep Reinforcement Learning," *IEEE Trans. Circuits Syst. II*, vol. 69, no. 3, pp. 1777–1781, Mar. 2022, doi: 10.1109/TCSII.2021.3107535.

[4] Y. Tang *et al.*, "Reinforcement Learning Based Efficiency Optimization Scheme for the DAB DC–DC Converter With Triple-Phase-Shift Modulation," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7350–7361, Aug. 2021, doi: 10.1109/TIE.2020.3007113.

[5] D. Jakobeit, M. Schenke, and O. Wallscheid, "Meta-Reinforcement-Learning-Based Current Control of Permanent Magnet Synchronous Motor Drives for a Wide Range of Power Classes," *IEEE Trans. Power Electron.*, vol. 38, no. 7, pp. 8062–8074, July 2023, doi: 10.1109/TPEL.2023.3256424.

[6] Q. Liu, Y. Guo, L. Deng, H. Liu, D. Li, and H. Sun, "Residual Deep Reinforcement Learning for Inverter-based Volt-Var Control," Aug. 13, 2024, *arXiv*: arXiv:2408.06790. Accessed: Nov. 16, 2024. [Online]. Available: http://arxiv.org/abs/2408.06790

[7] J. Ye, H. Guo, B. Wang, and X. Zhang, "Deep Deterministic Policy Gradient Algorithm Based Reinforcement Learning Controller for Single-Inductor Multiple-Output DC–DC Converter," *IEEE Trans. Power Electron.*, vol. 39, no. 4, pp. 4078–4090, Apr. 2024, doi: 10.1109/TPEL.2024.3350181.

[8] M. Dong, R. Liang, J. Yang, C. Xu, D. Song, and J. Wan, "Topology Derivation of Multiport DC–DC Converters Based on Reinforcement Learning," *IEEE Trans. Power Electron.*, vol. 38, no. 4, pp. 5055–5064, Apr. 2023, doi: 10.1109/TPEL.2023.3235053.

[9] D. Lee, B. Kim, S. Kwon, N.-D. Nguyen, M. Kyu Sim, and Y. Il Lee, "Reinforcement Learning-Based Control of DC-DC Buck Converter Considering Controller Time Delay," *IEEE Access*, vol. 12, pp. 118442–118452, 2024, doi: 10.1109/ACCESS.2024.3448535.

[10] A. Rajamallaiah, S. P. K. Karri, and Y. R. Shankar, "Deep Reinforcement Learning Based Control Strategy for Voltage Regulation of DC-DC Buck Converter Feeding CPLs in DC Microgrid," *IEEE Access*, vol. 12, pp. 17419–17430, 2024, doi: 10.1109/ACCESS.2024.3358412.

[11] N. Mazaheri, D. Santamargarita, E. Bueno, D. Pizarro, and S. Cobreces, "A Deep Reinforcement Learning Approach to DC-DC Power Electronic Converter Control with Practical Considerations," *Energies*, vol. 17, no. 14, p. 3578, July 2024, doi: 10.3390/en17143578.

[12] D. Alfred, D. Czarkowski, and J. Teng, "Reinforcement Learning-Based Control of a Power Electronic Converter," *Mathematics*, vol. 12, no. 5, p. 671, Feb. 2024, doi: 10.3390/math12050671.

[13] P. Qashqai, M. Babaie, R. Zgheib, and K. Al-Haddad, "A Model-Free Switching and Control Method for Three-Level Neutral Point Clamped Converter Using Deep Reinforcement Learning," *IEEE Access*, vol. 11, pp. 105394–105409, 2023, doi: 10.1109/ACCESS.2023.3318264.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," Aug. 28, 2017, *arXiv*: arXiv:1707.06347. doi: 10.48550/arXiv.1707.06347.

NOBORU KATAYAMA (M'11) received the B.S., M.S., and Ph.D. degrees in engineering from Tokyo University of Science, Japan, in 2006, 2008, and 2011, respectively. He is currently an Associate Professor in the Department of Electrical Engineering, Faculty of Science and Technology, Tokyo University of Science from 2020. His research interests include hydrogen energy, energy device diagnosis, and energy management. Dr. Katayama is a Member of the Institute of Electrical Engineers of Japan, and IEEE