IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 24 August 2023, accepted 14 September 2023, date of publication 22 September 2023,
date of current version 29 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318264

RESEARCH ARTICLE

A Model-Free Switching and Control Method for Three-Level Neutral Point Clamped Converter Using Deep Reinforcement Learning

POURIA QASHQAI[1], (Graduate Student Member, IEEE),
MOHAMMAD BABAIE[1], (Graduate Student Member, IEEE),
RAWAD ZGHEIB[2], (Member, IEEE), AND
KAMAL AL-HADDAD[1], (Life Fellow, IEEE)
[1]Department of Electrical Engineering, École de technologie supérieure (ÉTS), Montreal, QC H3C 1K3, Canada
[2]Institut de recherche d'Hydro-Québec (IREQ), Varennes, QC J3X 1S1, Canada

Corresponding author: Pouria Qashqai (pouria.qashqai@gmail.com)

This work was supported in part by OPAL-RT Technologies, QC, Canada; and in part by Institut de recherche d'Hydro-Québec (IREQ), QC, Canada.

ABSTRACT This paper presents a novel model-free switching and control method for three-level neutral point clamped (NPC) converter using deep reinforcement learning (DRL). Our approach targets two primary control objectives: voltage balancing and current control. In this method, voltage balancing, current control and selection of optimal switches are achieved using a reward function which is calculated based on various signals measured as observations of the DRL agent. Since the action space is discrete, a deep Q-network (DQN) agent is utilized. DQN is used due to its capability of handling high-dimensional state spaces. In order to highlight its pros and cons, the proposed method is compared with model predictive control (MPC), which is another popular non-linear control method for power electronic converters. The proposed method is evaluated and compared with the MPC method in grid-connected mode using simulations in Matlab/Simulink. To evaluate the practical performance of the DRL method, various experimental results are obtained. The simulation and experimental results demonstrate that the proposed method effectively achieves accurate voltage balancing and ensures steady operation even in the presence of various dynamic changes, including variations in the reference currents and grid voltage. Additionally, the method successfully handles uncertainties, such as sensor measurement noise, and accommodates parameter variations, such as changes in the capacity of the DC-link capacitors and line impedance. The results demonstrate that this method exhibits superior adaptability to real-time changes and uncertainties, delivering more robust performance compared to similar conventional methods like MPC. Thus, this method can be considered a promising approach for intelligent control of power electronic converters, especially when conventional methods such as MPC face challenges in performance and accuracy under severe parameter variations and uncertainties.

INDEX TERMS Artificial intelligence, deep reinforcement learning, machine learning, neutral point clamped, power converters, power electronics converters.

I. INTRODUCTION

Power electronics converters are essential components of numerous modern applications such as power supplies, renewable energies, electrical vehicles, and energy storage units [1], [2]. Control of power electronics converters is crucial to ensure their reliable, consistent, and efficient performance [3].

Linear control methods for power electronics converters, such as PID, are based on linear approximations of the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhilei Yao.

converters using linearization techniques such as averaging and state space modeling [4]. These methods are easy to study and implement while providing sufficient performance for simpler converters with linear or weakly non-linear operating modes. However, due to the small-signal nature of these methods, they are not proper for highly non-linear operating modes, uncertainties, and disturbances [5]. Some solutions are proposed in the literature to mitigate these drawbacks using several controllers to tackle different operating modes [6].

On the other hand, non-linear control methods use non-linear feedback or optimization techniques to address the non-linearity in power electronics converters and consequently provide robust performance across different operating modes. There are various non-linear control methods proposed in the literature but some of the most popular methods are feedback linearization, sliding mode control (SMC), and model predictive control (MPC).

Feedback linearization [7] transforms the non-linearity of converters into a linear representation so that the outcome could be controlled using conventional linear methods. This method requires an accurate model of the converter as well as its inverse model which may be difficult to extract in some complex converters. It is difficult to model unknown internal factors such as parasitic elements using this method which may lead to degradation in its performance. This method may be sensitive to sudden and large disturbances as well as noise in measurement signals.

Sliding mode control (SMC) [8], [9] forces the converter to reach and stay on a predetermined sliding surface. This method is resilient to uncertainties, external disturbances, and internal parameter variations. However, it introduces high switching frequencies to the switches which may result in switching loss, shorter lifespan of the switches, and efficiency degradation. Additionally, it requires a high-gain controller which makes this method sensitive to noise in measurements.

Model predictive control (MPC) [10], [11], [12] is inherently an online optimization problem that utilizes an accurate model of the converter to predict its future behavior. Then, it will be able to choose the best sets of actions from all the possible ones to aim for the highest possible optimization. Although this method can address non-linearities and constraints, it requires an accurate model of the converter and load, which may be difficult in practice. Moreover, this method demands high computational power and fast processing for solving an optimization problem at every single time stamp. Finally, since it relies on accurate modeling of the converter, this method is sensitive to parameter variations and other changes in system dynamics [13], [14]. Some studies have taken advantage of machine learning for tuning MPC parameters. For instance, in [15] a supervised learning model predictive control (SLMPC) method is utilized and optimized using the artificial bee colony (ABC) algorithm, to train the weighting factors of the cost function for controlling a three-phase NPC. This method replaces the conventional time-consuming and imprecise methods.

On top of all the challenges the aforementioned that conventional non-linear control methods are facing, they also suffer from performance degradation due to the accumulative error between a model and the actual behavior of a given power electronics converter [16], [17]. Various intelligent methods such as particle swarm optimization (PSO) and fuzzy neural network (FNN) are used alone or in conjunction with each other to mitigate these problems [18], [19]. However, these methods often suffer from weak adaptability and limited learning capacity. For instance, Control methods combined with PSO suffer from slow convergence and local optimal [20], whereas FNN has high computational complexity and overfitting problems [21]. Thus, more advanced non-linear control methods have gained popularity in recent years to mitigate these problems.

DRL is a subset of machine learning (ML) that utilizes deep neural networks (DNN) in reinforcement learning (RL) so that an agent learns from interactions with an environment to achieve the maximum long-term reward. In recent years, due to the breakthroughs in deep learning and the availability of high computational power, Deep reinforcement learning (DRL) has gained popularity to solve various complex problems including but not limited to robotics [22], [23], electrical vehicles and hybrid vehicles [24], [25], renewable energies [26], [27] and power systems [27], [28].

The advantages and disadvantages of the aforementioned methods are listed in Table 1. As seen, despite requiring high computational power, long training times, and large training datasets, DRL is utilized as an advanced control method for power electronic converters due to its resilience towards uncertainties, noise in measurements, and parameter variations.

TABLE 1. Overview of non-linear control methods for power electronic converters.

Control Method	Advantage(s)	Disadvantage(s)
Feedback Linearization	1. Handles non-linearity 2. Uses linear control techniques.	1. Needs precise models 2. Sensitive to disturbances
Sliding Mode Control (SMC)	1. Resilient to disturbances 2. Robust to model inaccuracies	1. High switching frequencies 2. Noise sensitive
Model Predictive Control (MPC)	1. Handles non-linearities and constraints 2. Predictive of future system behavior	1. Requires high computational power 2. Sensitive to parameter variations
Deep Reinforcement Learning (DRL)	1. Adaptable to complex tasks 2. Model-free	1. High computational needs 2. Needs large, diverse training data

In [29], DRL is applied as a solution to the shortcoming of conventional methods for control of DC/DC buck converter when feeding constant power load (CPL). The conventional methods often demonstrate poor performance in the presence of large changes in the CPL. Although the DRL method proposed in this paper can deliver satisfactory performance, the agent is used for tuning gains in the feedback loop. Thus, not only auxiliary control is required, but also a significant advantage of DRL which is being model-free is not exploited.

In [30], a model-free DRL controlling method for DC/DC buck converter feeding CPL is proposed. This method demonstrates good dynamic performance when large changes are applied to the CPL. However, the controller proposed in this paper requires accurate measurements which results in sensitivity to noise and accumulative error between a trained model and a real-world converter. Thus, this method may not be able to provide reliable performance in practice.

In [31], a novel technique is used to enable a DRL-controlled DC/DC buck converter to demonstrate resilient performance facing uncertainties and parameter variations. However, this method requires an off-line pre-trained model of the converter so that an extended state observer (ESO) observes the error between this model and the real-world converter to adapt to parameter variations and uncertainties.

In recent years, due to the significant breakthroughs in artificial intelligent algorithms especially machine learning (ML), and thanks to the remarkable surge in computational capabilities, research on the use of artificial intelligence in power electronics has gained substantial popularity [32]. Some studies have focused on machine learning-based modeling approaches [33], [34], [35], [36], while others have utilized machine learning to enhance the control of power electronics.

Although a considerable amount of research is focused on applications of DRL in power electronics, the status of research on the advantages of this method for the control of power electronic converters is still in its infancy [32]. For instance, In [37], DRL is used for control of a simple buck converter but no experimental tests are performed for evaluation. Similarly in [38], a buck converter is controlled by DRL and real-time simulations are performed for evaluation but no real-life practical results are obtained. In [39], a hybrid method is implemented for the control of a two-level converter. A DRL agent is utilized for obtaining the weighting factor design of an MPC controller. Few papers have studied applications of DRL in more complex converters such as Neutral Point Clamped (NPC) converters. For instance, in [40], a DRL agent is used for the efficiency optimization design of a three-level NPC. In [41], a model-free DRL method for controlling the three-level NPC is proposed. This method utilizes an actor-critic method to apply all the possible switching states under different conditions to learn the optimal switching algorithm that satisfies the control objectives. Although this method demonstrates a satisfactory steady-state performance, it is not studied under dynamic changes, uncertainties, internal and external parameter variations, and noise in measurements. Additionally, experimental results are not obtained to prove the practical feasibility of the proposed control method. Most importantly, the proposed method is not compared against a conventional non-linear control method like MPC.

As seen, despite the recent gaining interest in this subject, there is a gap in research on the application of DRL as a control method for multilevel power electronics converters such as NPC, emphasizing its advantages regarding uncertainties which conventional non-linear methods face challenges. This paper aims at improving the literature on this subject by proposing a new control method using DRL.

In this paper, a model-free DRL control method for three-level NPC is proposed. Through simulation results as well as experimental results, the method is proven to be resilient facing dynamic changes, uncertainties, internal and external parameter variations, and noise in measurements without requiring any auxiliary controllers or pre-trained models. Thus, this method may be a promising solution to the common problems associated with conventional non-linear control methods for power electronic converters.

The rest of this paper is structured as follows: In Section II, the fundamentals of DRL, various popular DRL methods as well as their advantages and disadvantages are explained in detail. In section III, the proposed method is introduced. Simulation results are obtained and discussed in section IV where the proposed method is compared with another conventional non-linear control method. Section V is dedicated to the experimental results that prove the real-world feasibility of the method. Finally, the limitations of the proposed method as well as the potential future studies are explained in section VI, and the whole paper is concluded.

II. FUNDAMENTALS OF DEEP REINFORCEMENT LEARNING

In this section, reinforcement learning will be introduced. Then different methods for finding optimal policy are explained and their limitations are discussed. Finally, a modern type of RL that utilizes deep learning (DL), also known as deep reinforcement learning (DRL) is introduced.

A. WHAT IS REINFORCEMENT LEARNING?

As shown in Fig. 1, machine learning (ML) is a subset of artificial intelligence (AI). Machine learning is a term used to describe various methods that enable computer programs to make decisions or predictions solely by learning from data. The major difference between these methods and numerical methods is that these methods are capable of generalization. There are three major categories of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning as depicted in Fig. 2.

Supervised learning is inherently an advanced curve-fitting method that learns the relationship between data and maps the input data and their corresponding target data. These sets of data can be labels (classification) or numbers (regression).

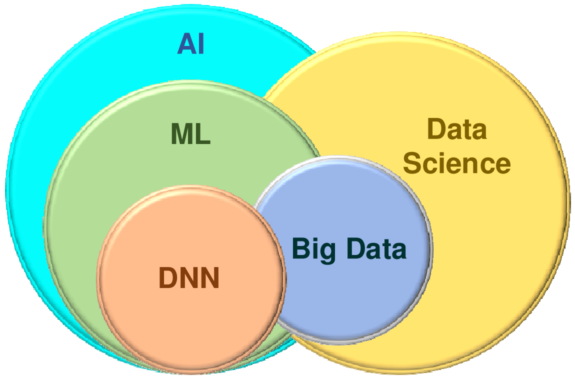


FIGURE 1. Diagram of overlap between data science and its subset big data with artificial intelligence (AI) and its subsets machine learning (ML) and deep neural networks (DNN).

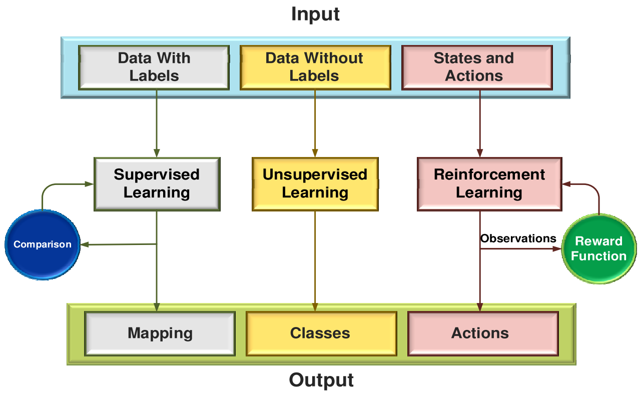


FIGURE 2. Diagram of different machine learning techniques.

Unsupervised learning, on the other hand, finds the relationship between data sets without using predetermined labels. This method is popular for finding anomalies in data and fault detection.

Reinforcement learning, as shown in Fig. 3., finds an optimal policy to achieve maximum cumulative reward through interactions with the environment without having access to the model of the environment. Learning solely by trial and error and learning from past experiences, enables RL to be used in various applications where extracting an accurate model is difficult or impossible.

As shown in Fig. 3, reinforcement learning is comprised of an agent, the environment, action, state, and reward. This can be expressed as a four-tuple Markov Decision Process (MDP) of $\{s, a, p_a, r_a\}$ where s is the current state, a is the action, p_a is the probability of going to state s by taking action a in state s and finally r_a_is the immediate reward. The objective of the RL agent is to find an optimal policy that leads to the maximum accumulative reward as shown in (1):

$$\pi^* = \operatorname{argmax}_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0, \pi\right] \tag{1}$$

π^* is an optimal policy continuously updated through (1), r_t is the reward at the time step t where the initial state is s_0, and γ is a discount factor that indicates how much the previous actions with their corresponding rewards can affect the future reward.

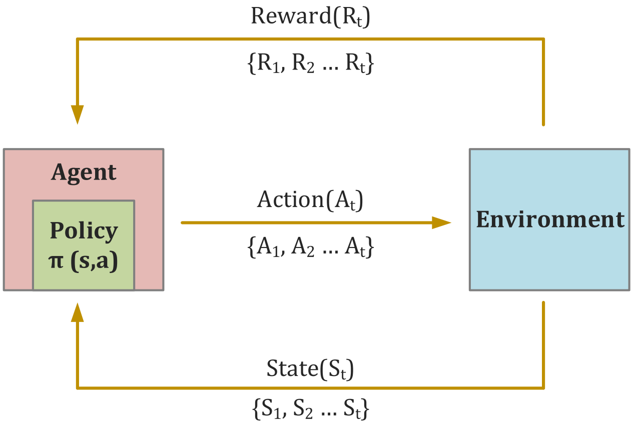


FIGURE 3. Diagram of a reinforcement learning agent exploring an environment, updating a policy of π (s,a) at time step of t.

B. FINDING THE OPTIMAL POLICY

There are three main methods for solving RL problems and obtaining the optimal policy for maximum accumulative long-term reward.

1) VALUE-BASED METHODS

These methods try to find the optimal value function without explicitly dealing with the policy. Using iterative algorithms such as Q-learning [42] and SARSA [43], these methods update a value function. This value function can predict the reward of taking action based on current states and previous rewards. To update the value function, the Bellman equation is utilized:

$$V(s, a) = R(s, a) + \gamma \max \left(V\left(s', a'\right)\right) \tag{2}$$

$V(s, a)$ is the value function in state s, $R(s, a)$ is the immediate reward for taking action a in state s, γ is the discount factor that determines how much the future rewards are taken into account, and finally $maxV\left(s'\right)$ is the maximum value function in the state s' for taking all actions of a'. Where a' represents the action that maximizes the value function in the state s'.

There are several disadvantages to using value-based methods, such as slow convergence rates, especially for high-dimensional state spaces, the inherent difficulty in handling continuous action spaces, and suboptimal performance in capturing time information in time series. To mitigate these problems, DQN-D and DRQN are proposed [44]. Despite these efforts, value-based methods are still recommended to be implemented in discrete action spaces.

2) POLICY-BASED METHODS

These methods try to find the optimal policy function without calculating the value function first. Using gradient-based algorithms such as Proximal Policy Optimization (PPO) [45] and Advantage Actor-Critic (A2C) [46], these methods update the policy parameters judging by their performance. In (3) the equation for updating the policy parameters in gradient ascent, one of the most popular methods is depicted.

$$\theta \leftarrow \theta + \propto \nabla_{\theta} J(\theta) \tag{3}$$

θ is the policy parameter, $\propto$ is the learning rate, $J(\theta)$ is the objective function, and finally $\nabla_{\theta} J(\theta)$ is the gradient of the objective function $J(\theta)$.

These methods suffer from several disadvantages such as high fluctuations in training episodes, becoming stuck in local optimal, and difficulty in handling discrete actions [47].

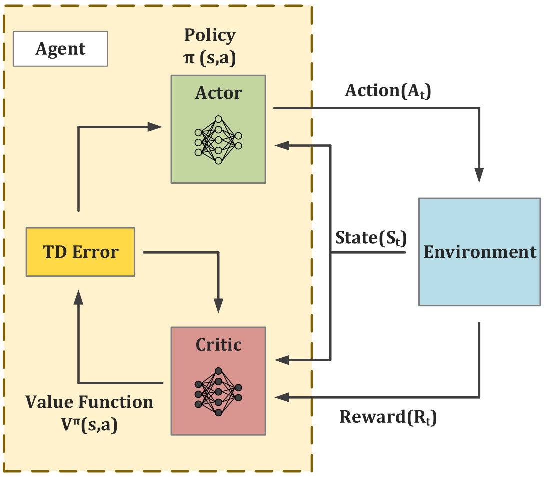


FIGURE 4. Block diagram of an actor-critic reinforcement learning method.

3) ACTOR-CRITIC METHODS

These methods combine the two previous methods in the form of two neural networks. The neural network which is responsible for policy is called ''actor'', whereas the neural network responsible for value function is called ''critic''. A diagram of an actor-critic RL is depicted in Fig. 4. As it can be seen, the RL agent uses the actor to generate actions and the critic evaluates how well the actions were chosen based on the value function. The actor judges by its temporal difference (TD) error. This enables the actor-critic methods to be more data-efficient compared to other methods.

However, these methods suffer from noisy gradients which leads to unstable learning and sensitivity to hyperparameters which leads to unreliable performance in disturbances [48].

C. DEEP REINFORCEMENT LEARNING

Conventional RL methods use linear function approximations. Therefore, their applications are limited to less complex problems with low levels of non-linearity. These methods may not be able to find an optimal policy when large state spaces or action spaces are present. By implementing neural networks into RL, as shown in Fig. 5, a more powerful generation of RL methods titled Deep Reinforcement Learning (DRL) emerged. Using neural networks provides several advantages in comparison to conventional RL methods: first, neural networks can be generalized, which makes them effective in dealing with uncertainties and unseen scenarios. Secondly, neural networks are capable of effectively handling high dimensions of input and output which enables them to deal with large state and action spaces. Moreover, neural networks can approximate non-linear functions, making them suitable for highly complex applications. And finally, neural networks can learn from experience, which enables them to learn a policy that is not known beforehand or may change in the future [49].

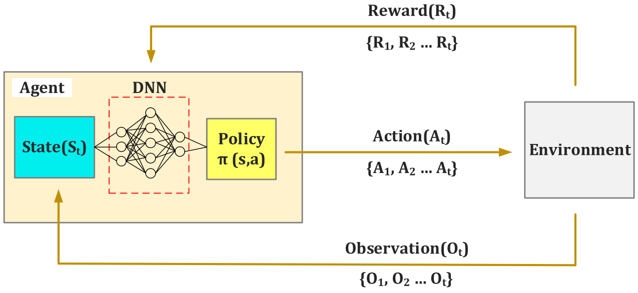


FIGURE 5. Block diagram of a reinforcement learning agent exploring an environment, updating a policy of π (s, a) at the time step of t.

III. THE PROPOSED METHOD

This section discusses the different aspects of the proposed method. First, the agent type is chosen. Secondly, observations that are signals measured by sensors, are selected. Then reward function and action space are proposed. Finally, the implementation of the DRL agent on a three-level NPC, as a switching and control method is discussed. The overall diagram of the proposed method is shown in Fig. 6.

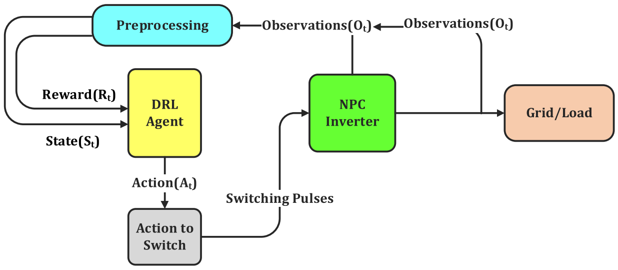


FIGURE 6. Block diagram of the proposed method using a DRL agent at time t to generate action A_t which is converted to switching pulses using the ''Action to Switch'' algorithm.

A. SELECTING AGENT TYPE

As mentioned in the previous section, there are different methods for finding the optimal policy for RL agents. Each of these methods has its advantages and disadvantages. Policy-based methods are more efficient when action space is continuous, however, value-based methods are better at handling large continuous state spaces while tackling discrete action spaces [47]. Actor-critic methods are effective in both continuous and discrete state spaces and action spaces.

TABLE 2. Popular agent types and other action spaces.

Agent	Type	Action Space	On/Off Policy
Q-Learning Agents (Q)	Value-Based	Discrete	Off
SARSA Agents	Value-Based	Discrete	On
Deep Q-Network (DQN) Agents	Value-Based	Discrete	Off
Policy Gradient Agents (PG)	Policy-Based	Discrete or continuous	On
Actor-Critic Agents (AC)	Actor-Critic	Discrete or continuous	On
Deep Deterministic Policy Gradient (DDPG) Agents	Actor-Critic	Continuous	Off
Twin-Delayed Deep Deterministic Policy Gradient Agents (TD3)	Actor-Critic	Continuous	Off
Soft Actor-Critic Agents (SAC)	Actor-Critic	Continuous	Off
Proximal Policy Optimization Agents (PPO)	Actor-Critic	Discrete or continuous	On
Trust Region Policy Optimization Agents (TRPO)	Actor-Critic	Discrete or continuous	On
Model-Based Policy Optimization Agents (MBPO)	Actor-Critic	Discrete or continuous	Off

However, these methods require more computational power and memory due to utilizing two separate networks. They also suffer from instability, delayed reward, and extended training times due to the correlation problems between actors and critics [46]. The most popular DRL agent types supported by MATLAB [50] are listed in Table 2.

Since control of a three-level NPC is inherently comprised of a large continuous state space and a small action space, a value-based method like Q-learning would be a proper choice.

Moreover, since deep learning provides many advantages over conventional numerical methods of RL, in this paper, the deep network form of Q-learning which is Deep Q-network (DQN) is selected. A diagram of the DQN network is shown in Fig. 7. Unlike Q-learning, which uses a Q-table as a lookup table to store pairs of state-action and their values, DQN uses neural networks and a replay buffer to achieve the same goal.

B. OBSERVATIONS AND PREPROCESSING

To generate a proper state space as depicted in (4), various observations are required.

$$\mathrm{s}_t = \mathrm{f}\left(o_t\right) \tag{4}$$

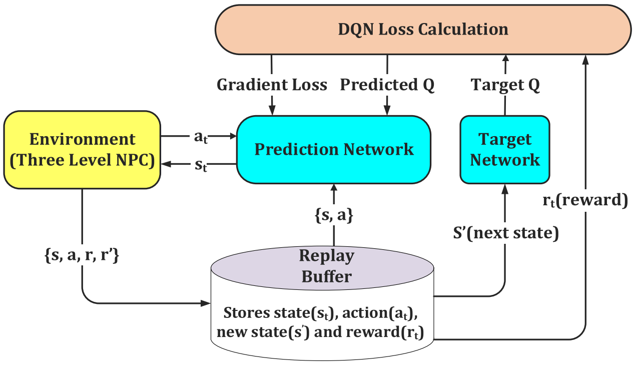


FIGURE 7. Topology of the deep Q-learning network (DQN) agent utilizing a replay buffer to find the optimal value function for control of the environment.

In (4), s_t is the state space at time t, whereas o_t is the matrix of observations at time t, and f is a function that manipulates the observations through a preprocessing algorithm to map them to the state space. In (5), the observation matrix is depicted:

$$o_t = \left[V_{abc}, id_{ref}, iq_{ref}, I_{abc}, V_{C1}, V_{DC}\right] \tag{5}$$

where V_{abc} and I_{abc} are the three-phase voltage and current of the grid (or load, depending on the operational mode) respectively, id_{ref} and iq_{ref} are the d and q reference currents in the park's transformation respectively, V_{C1} is the voltage across C_1 one of the capacitors in the DC link (since the voltage of C_2 is dependent on this voltage and balancing one capacitor is enough to balance both of them), and finally, V_{DC} is the voltage of the DC link.

Using the measurements in o_t may not converge or may lead to significantly high training times. Thus, a pre-processing unit is used to manipulate the signals so that the agent can determine the performance of its policy more effectively. The pre-processing block is shown in Fig. 8. As seen, the state space is continuous and comprised of various signals and measurements mapped to represent the behavior of the converter.

C. REWARD FUNCTION

Reward function creation is probably the most important and most challenging part of DRL. Since the agent has no agency, it can easily exploit undesirable outcomes to achieve an optimal reward. Therefore, it is vital to create a reward function that eliminates any potentially undesirable or non-sensical scenarios that may lead to high rewards. For instance, let us assume that a DRL agent is utilized for driving a vehicle and a reward function delivers a medium reward for maintaining speed and a high reward for not colliding with the environment. By constructing such a flawed reward function, the agent may decide not to move the vehicle at all to achieve maximum reward, which would be undesirable. Having this notion in mind, the following reward function is constructed

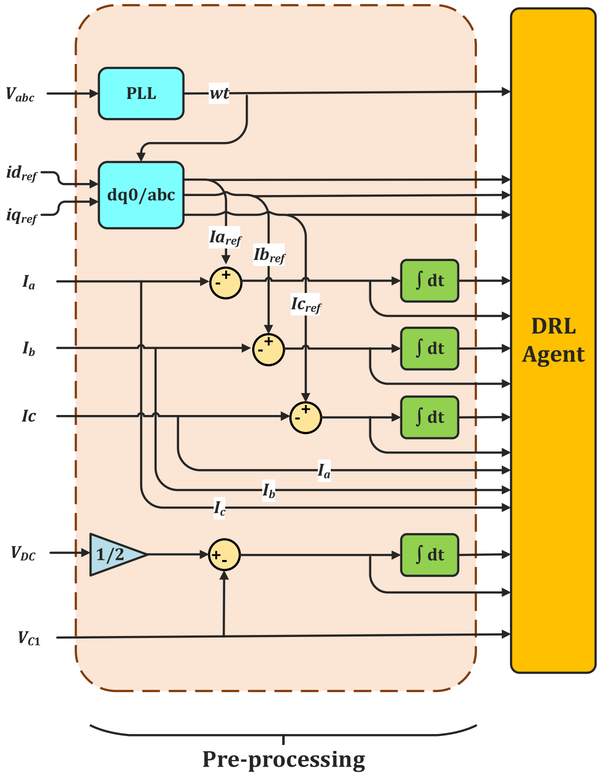


FIGURE 8. Diagram of the pre-processing unit that manipulates observations to become states that are comprehensible by the agent.

for this control algorithm:

$$R_T^t = \alpha(\varphi^{\mathrm{t}}(\mathrm{i_d}) + \varphi^{\mathrm{t}}\left(\mathrm{i_q}\right) + \varphi^{\mathrm{t}}(\mathrm{V_{C1}})) \tag{6}$$

where t is the time at any given time step. $\varphi(\mathrm{i_d})$ and $\varphi\left(i_q\right)$ are the reward functions of the d and q currents in park's transform respectively, $\varphi(V_{C1})$ is the reward function of the voltage across C_1, α is the gain, and finally, R_T is the total reward applied to the DQN agent. The reward functions of i_d, i_q and V_{C1} can be obtained using (7)-(9), respectively.

$$\varphi^{\mathrm{t}}(\mathrm{i_d}) = \begin{cases} -2\left|\nabla(i_d)\right|^2 & \left|\nabla(i_d)\right| > 0.2\text{ A} \\ -2\left|\nabla(i_d)\right|^2 + 2e^{-2} & \left|\nabla(i_d)\right| \leq 0.2\text{ A} \end{cases} \tag{7}$$

$$\varphi^{\mathrm{t}}\left(\mathrm{i_q}\right) = \begin{cases} -2\left|\nabla\left(i_q\right)\right|^2 & \left|\nabla\left(i_q\right)\right| > 0.2\text{ A} \\ -2\left|\nabla(q)\right|^2 + 2e^{-2} & \left|\nabla\left(i_q\right)\right| \leq 0.2\text{ A} \end{cases} \tag{8}$$

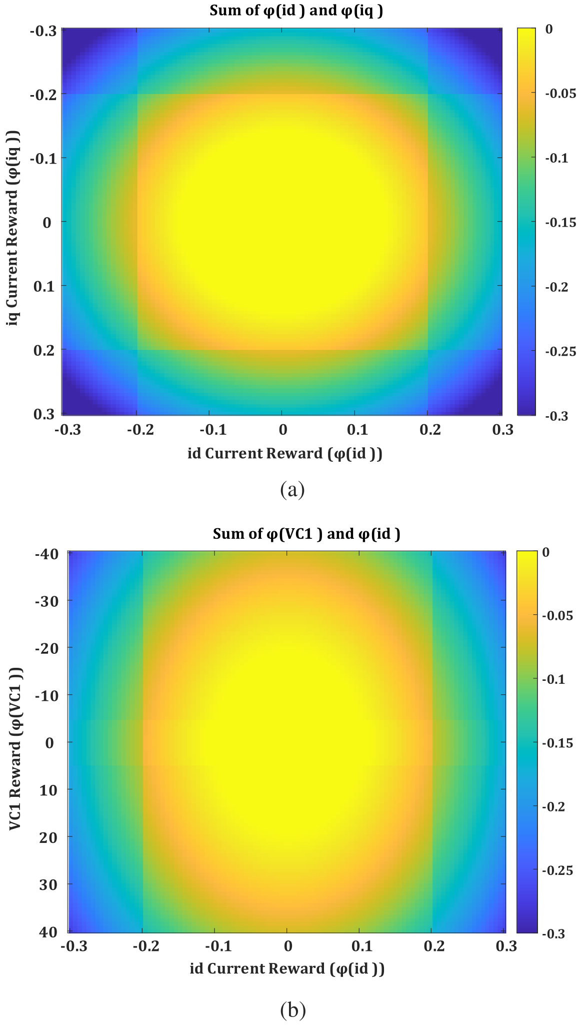
$$\varphi^{\mathrm{t}}(V_{C1}) = \begin{cases} -\left|\nabla(V_{C1})\right|^2 & \left|\nabla(V_{C1})\right| > 5\text{V} \\ -\left|\nabla(V_{C1})\right|^2 + 5e^{-3} & \left|\nabla(V_{C1})\right| \leq 5V \end{cases} \tag{9}$$

where $\nabla(i_d)$, $\nabla(i_q)$, and $\nabla(V_{C1})$ are the errors between each of these signals and their reference values, as shown in (10)-(12), respectively.

$$\nabla(i_d) = i_d - i_{d_{ref}} \tag{10}$$

$$\nabla\left(i_q\right) = i_q - i_{q_{ref}} \tag{11}$$

$$\nabla(i_d) = V_{C1} - \frac{V_{DC}}{2} \tag{12}$$



(a)

(b)

FIGURE 9. Heatmap of the combined reward for errors between i_d and i_q with their reference values (a) ; and errors between i_d and V_{C1} and their reference values (b).

To better understand the reward function, a heatmap of the combined rewards of $\varphi^t(i_d)$ and $\varphi^t\left(i_q\right)$, as well as, $\varphi^t(i_d)$ and $\varphi^t(V_{C1})$ are depicted in Fig. 9-a. and Fig. 9-b, respectively. In order to demonstrate the overall reward, a 3D graph of the total reward under three scenarios is depicted in Fig. 10. Scenario-1 is when only $\varphi^t(i_d)$ is considered. Scenario-2 combines the rewards of $\varphi^t\left(i_q\right)$ and $\varphi^t\left(i_q\right)$. Eventually, scenario-3, adds $\varphi^t\left(i_q\right)$ to the equation which makes it equal to the total reward. The x-axis represents the percentage of deviation from nominal errors of i_d, i_q, and V_{C1}. As can be seen, negative rewards are used to punish undesirable behaviors of the agent. Thus, the optimal reward for this agent is nearly zero.

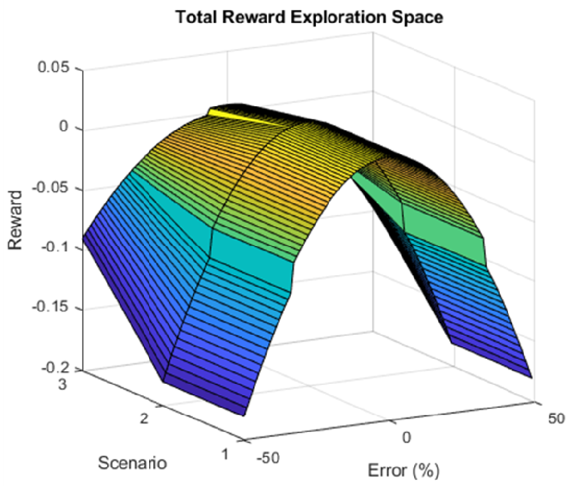


FIGURE 10. Exploration space for the DRL agent to achieve maximum reward.

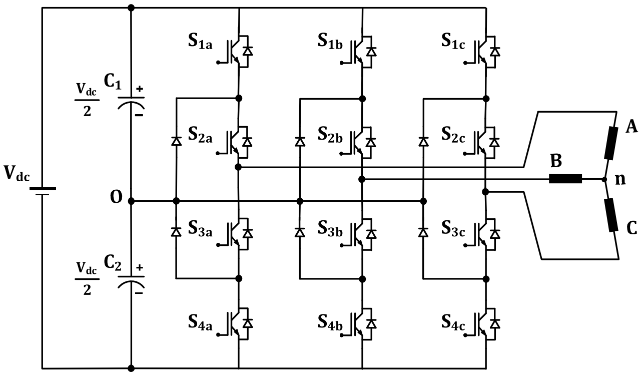


FIGURE 11. Topology of a three-level neutral point clamped (NPC) converter.

TABLE 3. Switching states of leg X of a three-level NPC.

Switching State	S_{1X}	S_{2X}	S_{3X}	S_{4X}	Output Voltage
P(+)	On	On	Off	Off	$+\frac{V_{dc}}{2}$
O	Off	On	On	Off	0
N(-)	Off	Off	On	On	$-\frac{V_{dc}}{2}$

D. ACTION SPACE

Before discussing the action space, consider the topology of a three-level NPC illustrated in Fig. 11. As seen, the converter is comprised of 12 switches. Since each switch has two states of On and Off. Therefore, there are $2^4 = 16$ combinations at each leg of the converter. However, only three combinations are permitted for each leg as is listed in Table 3 for leg X (i.e., A, B, or C). The rest of the combinations lead to fault or short-circuit. Since the converter has three permitted states and it has three legs, the total number of permitted switching states is $3^3 = 27$. Since these labels are unintelligible for the DRL agent, an integer number is used to represent each combination. Thus, the discrete action space in this method would be:

$$a_t = [0, 1, \ldots, 26] \tag{13}$$

Using a lookup table connected to the output of the DRL agent, each action which is an integer, is converted to its corresponding switching signals. As shown in Fig. 6, the "Action to Switch" block maps actions to switching signals.

E. CONNECTING THE AGENT TO THE CONVERTER

The diagram of a three-level NPC controlled by the proposed method is shown in Fig. 12. The actions taken by the DRL agent do not change until the next sampling period of the agent which is not necessarily equal to the sampling time of the simulation. Therefore, the sampling period of the DRL agent is equal to the switching frequency (f_{sw}).

IV. SIMULATION RESULTS

To evaluate the performance of the proposed method, we implemented it in Matlab/Simulink simulation environment usi. Simulation parameters, as well as training parameters, are listed in Table 4 and Table 5, respectively.

After training the DRL agent with a sampling time of 50μs, which is equal to a switching frequency of 20 KHz, the saved agent is initiated with a sampling time of 100 μs which is equal to a switching frequency of 10 kHz.

A. STEADY-STATE OPERATION

By setting the reference currents of i_{dref} and i_{qref} to 20 A and 0A, respectively, the steady-state results are obtained. The steady-state waveforms of the output i_d and i_q currents, output three-level currents, the voltage across the C_1 capacitor, and V_{an} the output phase voltage is illustrated in Fig. 13-a, Fig. 13-b, Fig. 13-c, Fig. 13-d, respectively. As can be seen, the DRL agent is capable of effectively following the reference currents, while balancing the dc-link capacitors. It is worth mentioning that the THD of the three-phase output currents is 3.62%, which is within the standard limits.

B. UNCERTAINTIES AND PARAMETER VARIATION

As mentioned earlier, the proposed method in this paper can take advantage of the characteristics of DRL to generalize its knowledge of the converter when facing uncertainties.

To evaluate the performance of the proposed method, the simulation is run for 300 milliseconds with different operation modes. Five scenarios are examined, and their simulation results are depicted in Fig. 14, where Fig. 14-a shows the output i_d and i_q currents, Fig. 14-b shows the output three-level currents, Fig. 14-c illustrates the voltage across the C_1 capacitor, and Fig. 14-d depicts V_{an} the output line voltage. As seen at time $t = 33ms$, i_{dref} is changed from 10A to 20A to evaluate the dynamic response of the DRL agent to sudden

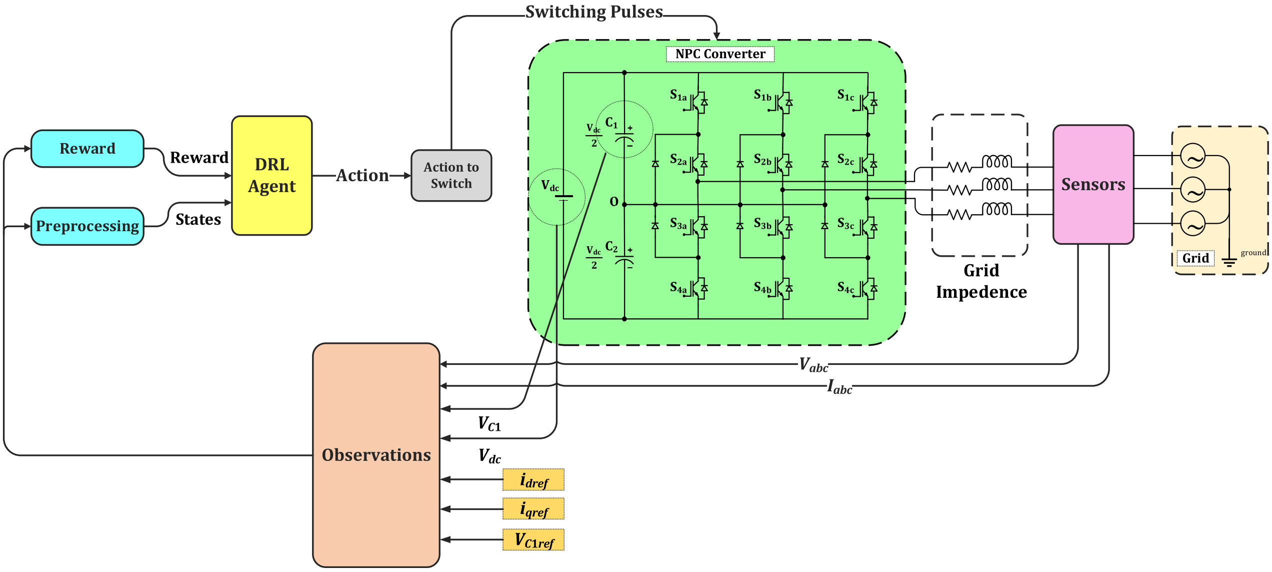


FIGURE 12. Block diagram of the proposed method for control and switching of a three-level NPC.

TABLE 4. Simulation parameters.

Parameter	Value	Unit
Grid Voltage (per phase)	170	V
Grid Frequency	60	Hz
Grid Resistance (per phase)	0.1	Ω
Grid Inductance (per phase)	5	mH
DC-link Capacitor(s)	1000	μF
DC-link Voltage	400	V

active power changes. Similarly, at time $t = 66ms$, i_{qref} is changed from zero to 20 A to evaluate the dynamic response of the DRL agent to sudden reactive power changes. To assess the resilience of the proposed method in the presence of external parameter variations, the grid inductance is increased by 20% at $t = 150ms$. There is a slight distortion in the waveforms but after a few samples, the agent adapts to the new condition. To evaluate the resilience of the proposed method in the presence of internal parameter variations, the capacity of each one of the dc-link capacitors is reduced by 15% at $t = 200ms$.

C. NOISE IN MEASUREMENT

Ultimately, to evaluate the robustness of the proposed method when there is noise in measurements, a white Gaussian noise (WGN) is added to each measurement signal at $t= 250ms$. The power of each added WGN is adjusted so that the signal-to-noise ratio (SNR) is 25 dB. As shown in (14) and (15), the root-mean-square (RMS) of the noise signal should be around 5% of the RMS of the measured signal so that SNR is approximately 25 dB. In (14) and (15), P and A are the power and RMS of signals or noise, respectively.

TABLE 5. Training parameters.

Parameter	Value
Layer Size of the State Path	140
Layer Size of the State Path	48
Learning Rate	0.001
Normalization	None
Bias Learn Rate Factor	0
Double DQN	No
Mini Batch Size	320
Discount Factor	0.01
Score Averaging Window Length	5
Agent Sampling Time	50µs

$$SNR = \frac{P_{signal}}{P_{noise}} = \left(\frac{A_{signal}}{A_{noise}}\right)^2 \tag{14}$$

$$SNR_{dB} = 10\log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{15}$$

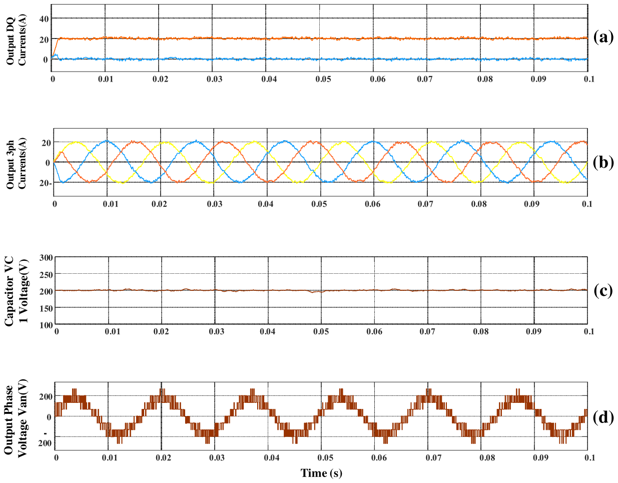
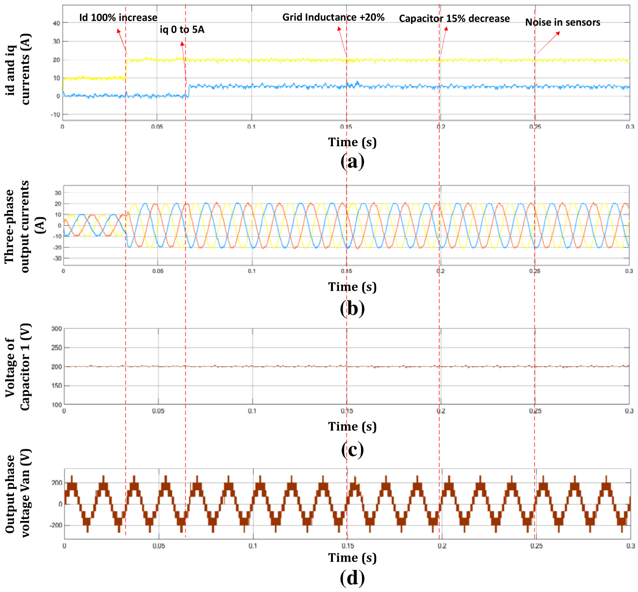


FIGURE 13. Waveforms of the output i_d and i_q currents (a) ; output three-level currents (b); the voltage across the C_1 capacitor (c); and V_{an} the output phase voltage (d), in steady-state operation, when $i_{dref} = 20$ and $i_q = 0$.



(a)

(b)

(c)

(d)

FIGURE 14. Waveforms of the output i_d and i_q currents (a) ; output three-phase currents (b); the voltage across the C_1 capacitor (c); and V_{an} the output phase voltage (d), when facing active power changes, reactive power changes, grid inductance increase, capacitor degradation, and noise in measurements.

By doing so it can be seen that the agent continues to perform satisfactorily despite the presence of noise in measurement.

D. COMPARISON WITH THE MODEL PREDICTIVE CONTROL (MPC) METHOD

To further evaluate the proposed method and assess its advantages, disadvantages, and limitations, the same topology and parameters are controlled by both the DRL method, and another conventional non-linear control method known as the Model Predictive Control (MPC). The cost function of the MPC controller is multi-objective as shown in (16).

$$G = \lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3 \tag{16}$$

It aims at controlling current, reducing common mode voltage (CMV) and capacitor voltages. Where $g_1 - g_3$ represent the reference current, capacitor voltages and CMV, respectively. Similarly, $\lambda_1 - \lambda_3$ are the corresponding weighting factors.

To perform a fair comparison, the DRL agent is re-trained using the same training options mentioned in Table 5. Only this time, the sampling time of both the DRL agent and the MPC controller is equally set to 50μs.

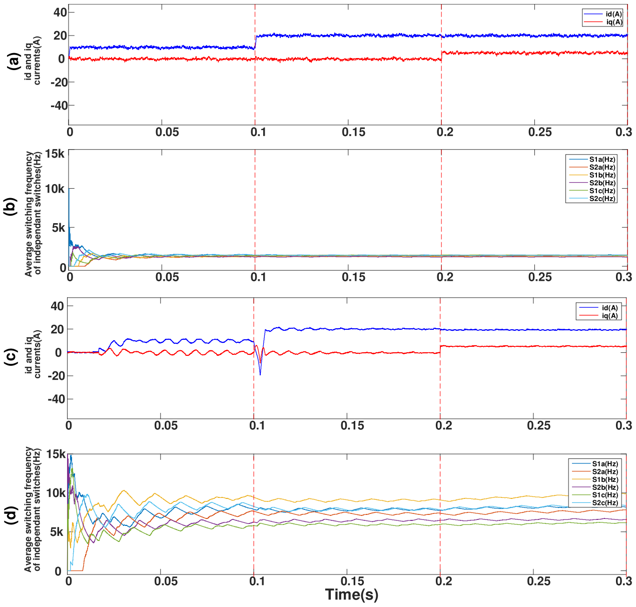


FIGURE 15. Waveforms of the output i_d and i_q currents using the DRL method (a) ; average frequency of the switches using the DRL method (b); the output i_d and i_q currents using the MPC method (c); and average frequency of the switches using the MPC method (d).

1) STEADY-STATE OPERATION AND STEP RESPONSE

To assess the performance of each method in terms of steady-state operation and step response, the reference current i_d is increased from 10 A to 20 A at $t = 100\ ms$ as shown in Fig. 15-a for the DRL agent and Fig. 15-c for the MPC controller, respectively. As seen, the initialization phase (warm-up) of the DRL method is superior to that of the MPC method. Additionally, the step response of the MPC controller results in a larger undershoot, more prolonged oscillations, and longer settling times. Similarly, as shown in Fig. 15-a and Fig. 15-c, the reference current of i_q is changed from 0 A to 5 A at $t = 200ms$. As depicted, both methods resulted in satisfactory transient results as well as accurate and smooth steady-state performances. It can be concluded that depending on the magnitude of step changes, the DRL agent demonstrates superior warm-up and transient responses compared to the MPC method.

2) SWITCHING LOSSES

In order to evaluate the effect of each method on switching losses the switching frequencies can be studied. By counting the rising edges of each switch and dividing it by time, the average switching frequencies corresponding to each switch can be calculated. Since in the NPC, there are supplementary switch couples where the state of each switch is the inverse of its dual switch (e.g., S_{1a} and S_{3a}), we have selected one switch from each couple. Thus, the switches S_{1a}, S_{2a}, S_{1b}, S_{2b}, S_{1c}, and S_{2c} are considered for an apparent presentation.

As you can see in Fig. 15-c and Fig. 15-d, despite having the same sampling time (T_S), the switching frequencies of switches in the DRL method are approximately around 2.5 KHz with a low degree of variance, whereas in the MPC method, not only the switching frequencies are higher and consequently the switching losses are higher but also the switching frequencies have a large degree of variance which makes the design of a filter more complex and more challenging.

In addition, considering the dynamic changes of the reference currents at $t = 100\ ms$ and $t = 200\ ms$ a depicted in Fig. 15-a and Fig. 15-c, we can conclude that both methods demonstrate near-constant switching frequencies when facing dynamic changes. However, in the MPC method, the average switching frequencies change slightly when the reference currents change but this change is negligible.

Ultimately, it is worth mentioning that despite having higher switching frequencies the MPC method results in slightly smoother current waveforms with a THD of 2.44% compared to the THD of the DRL method which equals 3.62%. Considering the less satisfactory performances of the NPC method in terms of switching frequency and switching losses, and despite resulting in slightly smoother waveforms and better THD, we can conclude that the DRL method is superior to the MPC method in this domain.

3) UNCERTAINTY AND PARAMETER VARIATIONS

As stated before, control methods based on ML including DRL have the ability to generalize policies to never-seen scenarios. Furthermore, the DRL method proposed in this paper is inherently model-free and consequently, not sensitive to parameter changes. In contrast, the MPC method which has an acceptable performance for nonlinear control of power electronic converters, not only requires the precise model of the converter but also cannot generalize its policy to unexplored areas and unknown scenarios. For this reason, we have performed a series of simulation tests in this section of the paper to evaluate and compare the resilience of each method when facing parameter changes and uncertainties.

Six scenarios are considered for comparison:

A step increase in the reference current i_d from 10 A to 20 A at $t = 100ms$ is applied to the DRL agent as shown in Fig. 16 and to the MPC controller as shown in Fig. 17, respectively. As can be seen and stated earlier, the transient response to this change is superior in the DRL method in terms of settling time, undershoot, and oscillations. In the MPC method the output phase voltage v_{ao} is slightly distorted. Although the MPC controller mitigates this distortion after a short time, such deformation is not present in the DRL

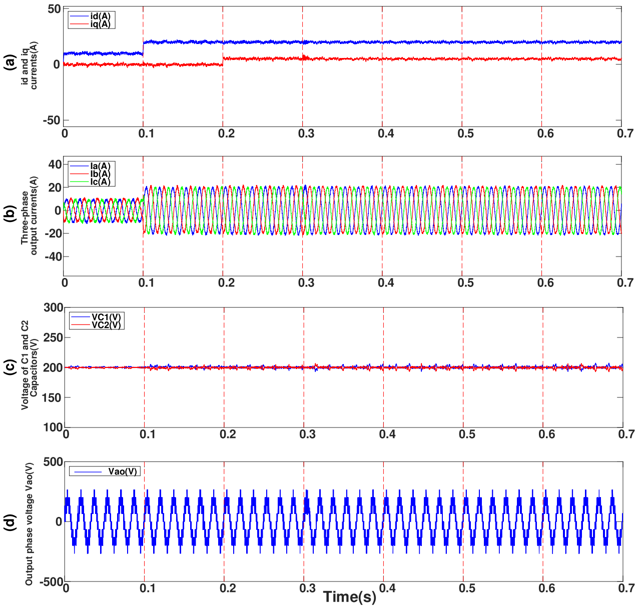


FIGURE 16. Waveforms of the output i_d and i_q currents (a) ; output three-phase currents (b); the voltage across the C_1 and C_2 capacitors (c); and V_{ao} the output phase voltage (d), when facing active power changes, reactive power changes, grid inductance increase, capacitor degradation, grid voltage increase and noise in measurements controlled by the DRL method.

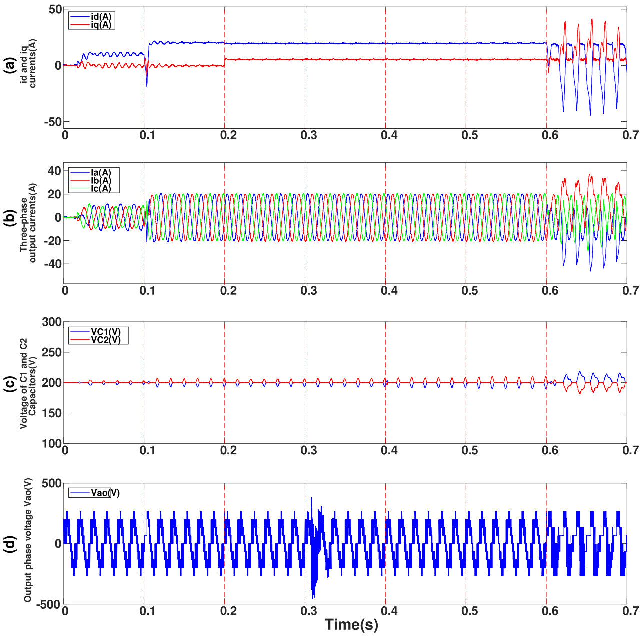


FIGURE 17. Waveforms of the output i_d and i_q currents (a) ; output three-phase currents (b); the voltage across the C_1 and C_2 capacitors (c); and V_{ao} the output phase voltage (d), when facing active power changes, reactive power changes, grid inductance increase, capacitor degradation, grid voltage increase and noise in measurements controlled by the MPC method.

method. In addition, the waveform of v_{ao} is more symmetrical in the DRL method. As stated before, THD is slightly lower in the MPC method resulting in smoother output current waveforms. Nonetheless, this smoothness in the waveform comes at the cost of higher switching frequencies and consequently higher switching losses. The voltages of the DC-link capacitors are successfully balanced in both methods, but their ripples are slightly increased in both methods upon increasing i_d.

A step increase in the reference current i_q from 0 A to 5 A at $t = 200ms$ is applied to the DRL method as shown in Fig. 16 and to the MPC method as shown in Fig. 17. The transient response of the DRL method is similar to the previous scenario while the transient response of the MPC method to a step change in i_q is better than the previous scenario. Contrasting the previous scenario, the waveform of v_{ao} has remained without distortion in both methods. The voltages of the DC-link capacitors remain unchanged.

The inductance and resistance of the grid line are increased by 20% at $t = 300ms$. This scenario is designed to evaluate the performance of both methods under grid line parameter variations. As seen in Fig. 16 and Fig. 17, this change has resulted in waveform distortion of the output current waveforms in both methods. This distortion of the waveforms is slightly worse in the DRL method. The waveform of the output phase voltage has distortion in both methods, but this distortion is mild in the DRL method but considerable in the MPC method. It is worth mentioning that these distortions are temporary and are eliminated after several sampling times. The waveforms of the voltages of the DC-link capacitors remain unchanged.

To simulate the performance under degraded capacitors, the capacitance of the DC-link capacitors is decreased by 15% at $t = 400ms$. As seen in Fig. 16 and Fig. 17, this change did not result in a considerable change in the output currents and the output phase voltage. However, the voltage ripples of the DC-link capacitors are increased which is not unexpected. Thus, it can be concluded that both methods demonstrate resilience toward moderate degradation in the DC-link capacitors.

The grid voltage is increased by 10% at $t = 500ms$. As seen in Fig. 16 and Fig. 17, the output currents of the converter and the output phase voltage v_{ao} remain unchanged. The voltages of the DC-link capacitors remain unchanged in the DRL method, but their ripples are increased slightly in the MPC method. It can be concluded that both methods demonstrate resilience toward variations of the voltage grid.

Ultimately, the most important scenario is considered in this section to evaluate the resistance of both methods when facing a common uncertainty which is noise in measurements. A noise of 25 dB is added to the measurements of the output currents I_{abc} at $t = 600ms$. As seen in Fig. 16 and Fig. 17 as soon as the noise is introduced the MPC controller becomes unstable. Not only do the output currents become severely distorted but also the output phase voltage v_{ao} is also heavily distorted and the voltage ripples of the DC-link capacitors are drastically increased. In the meantime, the DRL method is not affected by this uncertainty. This is due to its characteristic of being model-free and being able to generalize which are the inherent characteristics of machine learning algorithms.

In conclusion, based on the simulation results obtained in this section it can be stated that the MPC method despite demonstrating satisfactory performance in many scenarios is not a reliable solution when moderate uncertainties are present in the control environment. This is due to the fact that the MPC method requires a precise model of the converter, and it assumes the measurements are ideally obtained with minimal noise or distortion. On the other hand, the DRL method demonstrates resilience toward various kinds of parameter changes and uncertainties, especially noise in measurements. Thus, the DRL method despite being more computationally intensive is the better choice for applications where a moderate degree of uncertainty is present.

TABLE 6. Experimental test parameters.

Parameter	Value
DC-link voltage	400 V
Grid frequency	60 Hz
Grid filter inductance and resistance	5 mH, 0.1 Ω
DC-link capacitors	650 µF
Hardware sampling time	20 µs
Linear load inductance	50 mH
Linear load resistance	40 Ω
Nonlinear load	80 Ω, 2200 µF

V. EXPERIMENTAL RESULTS

In this section, the proposed DRL method has been experimentally evaluated using an advanced testbed constructed based on dSPACE 1202, OP8662 (voltage and current measurements), a power board of NPC, an autotransformer, and several loads detailed in Table 6. The proposed experimental setup is illustrated in Fig. 18, in which the test equipment is highlighted. The intelligent controller has been trained and implemented using the ‘‘Reinforcement Learning Toolbox’’ provided by MATLAB.

The steady-state performance of the NPC converter in the grid-connected mode is obtained when i_{dref} is set to 12 A and i_{qref} is 0 A, as depicted in Fig. 19. As seen, the agent effectively controls the NPC despite not having access to the mathematical model. The dynamic performance of

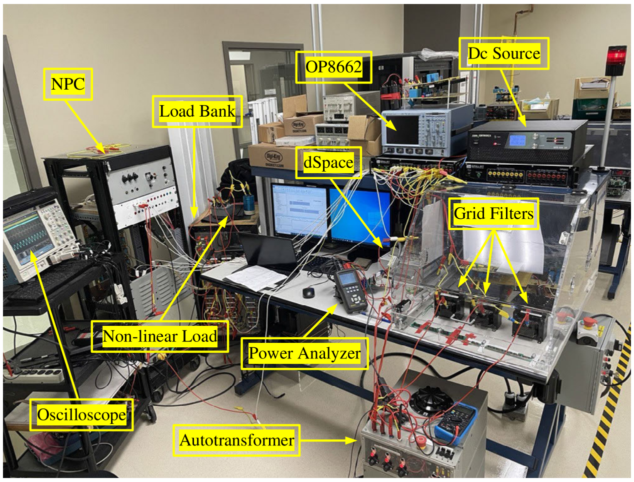


FIGURE 18. Details of the experimental setup built in the GREPCI laboratory to implement and evaluate the DRL control algorithm.

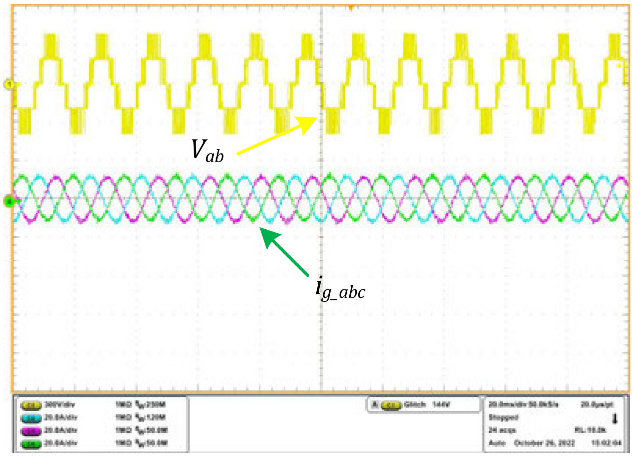


FIGURE 19. Experimental results of the line voltage and the three-phase current in steady-state mode when i_{dref} is set to 12A and i_{qref} is set to 0A.

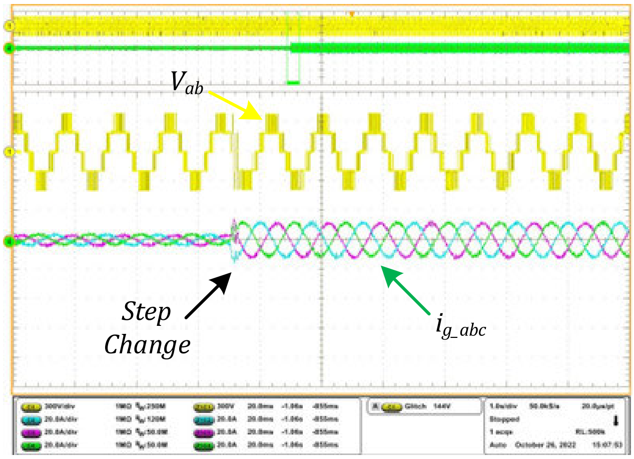


FIGURE 20. Experimental results of the line voltage and the three-phase currents under a step change of i_{dref} from 4A to 12A.

the proposed DRL controller under a step change of the d-reference current from 4 A to 12 A is shown in Fig. 20.

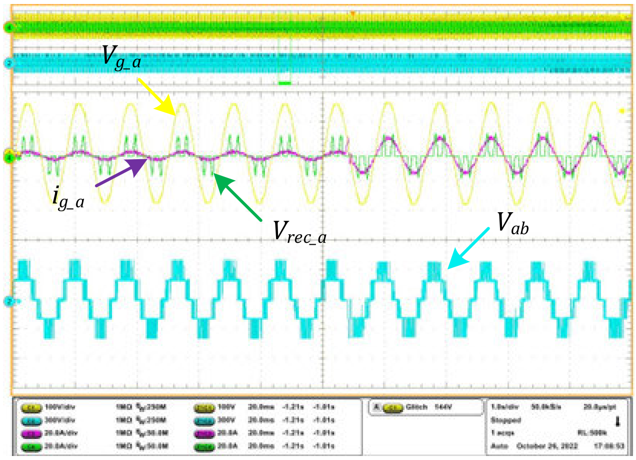


FIGURE 21. Dynamic performance of the proposed DRL controller under perturbations caused by a nonlinear load.

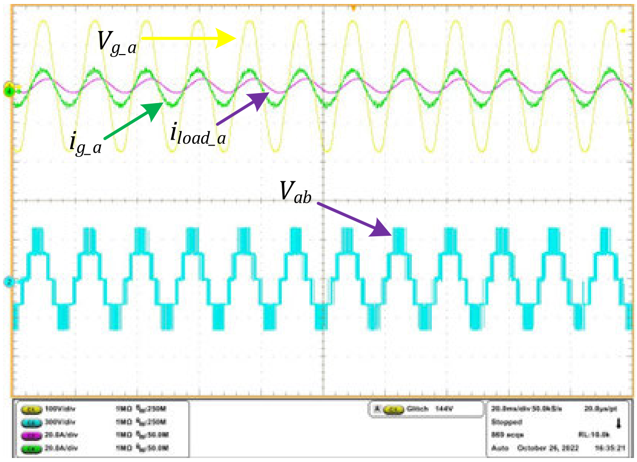


FIGURE 22. Experimental results of reactive power compensation under the operation of a reactive load.

The captured results from this dynamic test demonstrate the fast dynamic performance of the intelligent controller.

Ultimately, to evaluate the performance of the proposed method in the presence of non-linearity, it is connected to a full-bridge non-linear load with a dc impedance of 80 Ω and 2200 μF. The dynamic response of the proposed method in this scenario is obtained when a step change of i_{dref} from 3 A to 12 A is applied, as depicted in Fig. 21. Regarding the results of this test scenario, the proposed method demonstrates the robust performance of the DRL controller in the presence of perturbations caused by the non-linear load. In addition, Fig. 21 proves that the controller can perfectly synchronize the converter's current with the grid voltage, regardless of perturbations caused by loads or the grid environment. The experimental results of another test scenario in Fig. 22 indicate that the proposed DRL algorithm can successfully preserve the unity power factor of the grid by injecting reactive power demanded by loads. The THD analysis of the grid voltage and current in Fig. 23 also validates the optimal

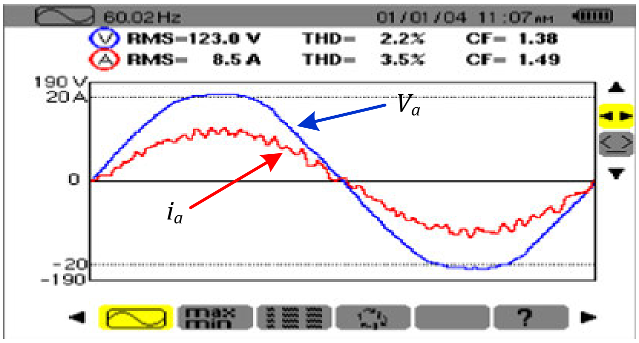


FIGURE 23. Experimental results of THD analysis of the grid voltage and current (phase A) under steady-state performance.

switching control performance of the proposed DRL control algorithm in practice.

VI. CONCLUSION

In conclusion, this paper presents a novel model-free switching and control method for a three-level NPC converter using deep reinforcement learning. The proposed method provides a robust performance under both simulation and experimental tests and achieves accurate voltage balancing and reference current tracking not only in steady-state mode but also under various dynamic changes, parameter variations, harmonic perturbations, and uncertainties. By applying various scenarios, the proposed method is compared against MPC, another conventional nonlinear control method for power electronic converters. The comparison results demonstrate the resilience of the proposed method against various types of parameter changes and uncertainties. The proposed DRL method is proven to be resilient against the presence of noise in sensor measurements whereas the MPC method becomes unstable in such an environment. However, it should be noted that this method requires long training times, and creating a reward function is very challenging. Nevertheless, future studies can be conducted to extend the application of this proposed method to other power electronic converters, especially those with complex control. Overall, this method can be considered a promising approach for model-free, non-linear control of power electronic converters, especially under parameter variations and uncertainties where conventional methods face performance and accuracy challenges.

REFERENCES

[1] G. Zhang, Z. Li, B. Zhang, and W. A. Halang, "Power electronics converters: Past, present and future," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 2028–2044, Jan. 2018.

[2] Y. P. Siwakoti, M. Forouzesh, and N. H. Pham, "Power electronics converters—An overview," in *Control of Power Electronic Converters and Systems*. New York, NY, USA: Academic Press2018, pp. 3–29.

[3] S. Bacha, I. Munteanu, and A. I. Bratcu, *Power Electronic Converters Modeling and Control* (Advanced Textbooks in Control and Signal Processing), vol. 454. London, U.K.: Springer, 2014, p. 454.

[4] A. Francés, R. Asensi, Ó. García, R. Prieto, and J. Uceda, "Modeling electronic power converters in smart DC microgrids—An overview," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6274–6287, Nov. 2018.

[5] Q. Xu, N. Vafamand, L. Chen, T. Dragicevic, L. Xie, and F. Blaabjerg, "Review on advanced control technologies for bidirectional DC/DC converters in DC microgrids," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 9, no. 2, pp. 1205–1221, Apr. 2021.

[6] A. Naziris, G. Guarderas, A. Francés, R. Asensi, and J. Uceda, "Large-signal black-box modelling of bidirectional battery charger for electric vehicles," in *Proc. IEEE Appl. Power Electron. Conf. Expo. (APEC)*, Mar. 2019, pp. 3195–3198, doi: 10.1109/APEC.2019.8721930.

[7] M. A. Mahmud, T. K. Roy, S. Saha, M. E. Haque, and H. R. Pota, "Robust nonlinear adaptive feedback linearizing decentralized controller design for islanded DC microgrids," *IEEE Trans. Ind. Appl.*, vol. 55, no. 5, pp. 5343–5352, Sep. 2019, doi: 10.1109/TIA.2019.2921028.

[8] H. Komurcugil, S. Biricik, S. Bayhan, and Z. Zhang, "Sliding mode control: Overview of its applications in power converters," *IEEE Ind. Electron. Mag.*, vol. 15, no. 1, pp. 40–49, Mar. 2021.

[9] F. Sebaaly, H. Vahedi, H. Y. Kanaan, N. Moubayed, and K. Al-Haddad, "Design and implementation of space vector modulation-based sliding mode control for grid-connected 3L-NPC inverter," *IEEE Trans. Ind. Electron.*, vol. 63, no. 12, pp. 7854–7863, Dec. 2016, doi: 10.1109/TIE.2016.2563381.

[10] J. Rodriguez et al., "Latest advances of model predictive control in electrical drives—Part I: Basic concepts and advanced strategies," *IEEE Trans. Power Electron.*, vol. 37, no. 4, pp. 3927–3942, Apr. 2022.

[11] J. Rodriguez et al., "Latest advances of model predictive control in electrical drives—Part II: Applications and benchmarking with classical control methods," *IEEE Trans. Power Electron.*, vol. 37, no. 5, pp. 5047–5061, May 2022.

[12] J. Hu, Y. Shan, J. M. Guerrero, A. Ioinovici, K. W. Chan, and J. Rodriguez, "Model predictive control of microgrids—An overview," *Renew. Sustain. Energy Rev.*, vol. 136, Feb. 2021, Art. no. 110422.

[13] M. Babaie, M. Mehrasa, M. Sharifzadeh, and K. Al-Haddad, "Floating weighting factors ANN-MPC based on Lyapunov stability for seven-level modified PUC active rectifier," *IEEE Trans. Ind. Electron.*, vol. 69, no. 1, pp. 387–398, Jan. 2022, doi: 10.1109/TIE.2021.3050375.

[14] M. Babaie and K. Al-Haddad, "ANN based model-free sliding mode control for grid-connected compact multilevel converters: An experimental validation," in *Proc. IEEE 30th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2021, pp. 1–6, doi: 10.1109/ISIE45552.2021.9576194.

[15] M. Babaie, M. Sharifzadeh, M. Mehrasa, G. Chouinard, and K. Al-Haddad, "Supervised learning model predictive control trained by ABC algorithm for common-mode voltage suppression in NPC inverter," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 9, no. 3, pp. 3446–3456, Jun. 2021.

[16] R. Darbali-Zamora and E. I. Ortiz-Rivera, "An overview into the effects of nonlinear phenomena in power electronic converters for photovoltaic applications," in *Proc. IEEE 46th Photovoltaic Specialists Conf. (PVSC)*, Jun. 2019, pp. 2908–2915, doi: 10.1109/PVSC40753.2019.8980933.

[17] K. F. Krommydas and A. T. Alexandridis, "Nonlinear analysis methods applied on grid-connected photovoltaic systems driven by power electronic converters," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 8, no. 4, pp. 3293–3306, Dec. 2020, doi: 10.1109/JESTPE.2020.2992969.

[18] H. S. Krishnamoorthy and T. N. Aayer, "Machine learning based modeling of power electronic converters," in *Proc. IEEE Energy Convers. Congr. Expo. (ECCE)*, Sep. 2019, pp. 666–672, doi: 10.1109/ECCE.2019.8912608.

[19] Y. Liao, Y. Li, M. Chen, L. Nordström, X. Wang, P. Mittal, and H. V. Poor, "Neural network design for impedance modeling of power electronic systems based on latent features," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 18, 2023, doi: 10.1109/TNNLS.2023.3235806.

[20] F.-Y. Huang, "A particle swarm optimized fuzzy neural network for credit risk evaluation," in *Proc. 2nd Int. Conf. Genetic Evol. Comput.*, Sep. 2008, pp. 153–157, doi: 10.1109/WGEC.2008.25.

[21] M. Wang and Y. Wang, "Fuzzy neural-network-based output tracking control for nonlinear systems with unknown dynamics," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2020, pp. 5124–5129, doi: 10.1109/CAC51589.2020.9327892.

[22] H. Nguyen and H. La, "Review of deep reinforcement learning for robot manipulation," in *Proc. 3rd IEEE Int. Conf. Robotic Comput. (IRC)*, Feb. 2019, pp. 590–595.

[23] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: Lessons we have learned," *Int. J. Robot. Res.*, vol. 40, nos. 4–5, pp. 698–721, Apr. 2021.

[24] G. Du, Y. Zou, X. Zhang, T. Liu, J. Wu, and D. He, "Deep reinforcement learning based energy management for a hybrid electric vehicle," *Energy*, vol. 201, Jun. 2020, Art. no. 117591.

[25] X. Hu, T. Liu, X. Qi, and M. Barth, "Reinforcement learning for hybrid and plug-in hybrid electric vehicle energy management: Recent advances and prospects," *IEEE Ind. Electron. Mag.*, vol. 13, no. 3, pp. 16–25, Sep. 2019.

[26] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Convers. Manage.*, vol. 198, Oct. 2019, Art. no. 111799.

[27] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, Mar. 2020.

[28] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Reinforcement learning and its applications in modern power and energy systems: A review," *J. Mod. Power Syst. Clean Energy*, vol. 8, no. 6, pp. 1029–1042, Nov. 2020, doi: 10.35833/MPCE.2020.000552.

[29] 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.

[30] T. Yang, C. Cui, and C. Zhang, "On the robustness enhancement of DRL controller for DC–DC converters in practical applications," in *Proc. IEEE 17th Int. Conf. Control Autom. (ICCA)*, Jun. 2022, pp. 225–230, doi: 10.1109/ICCA54724.2022.9831887.

[31] 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, Exp. Briefs*, vol. 69, no. 3, pp. 1777–1781, Mar. 2022, doi: 10.1109/TCSII.2021.3107535.

[32] P. Qashqai, H. Vahedi, and K. Al-Haddad, "Applications of artifical intelligence in power electronics," in *Proc. IEEE 28th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2019, pp. 764–769.

[33] P. Qashqai, K. Al-Haddad, and R. Zgheib, "Modeling power electronic converters using a method based on long-short term memory (LSTM) networks," in *Proc. IECON 46th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2020, pp. 4697–4702.

[34] P. Qashqai, R. Zgheib, and K. Al-Haddad, "GRU and LSTM comparison for black-box modeling of power electronic converters," in *Proc. 47th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2021, pp. 1–5.

[35] P. Qashqai, K. Al-Haddad, and R. Zgheib, "Deep neural network-based black-box modeling of power electronic converters using transfer learning," in *Proc. IEEE Energy Convers. Congr. Expo. (ECCE)*, Oct. 2022, pp. 1–6.

[36] P. Qashqai, R. Zgheib, and K. Al-Haddad, "A programmatical method for real-time simulation of black-box LSTM-based models of power electronic converters in Hypersim," in *Proc. IEEE 1st Ind. Electron. Soc. Annu. On-Line Conf. (ONCON)*, Dec. 2022, pp. 1–5.

[37] D. Alfred, D. Czarkowski, and J. Teng, "Model-free reinforcement-learning-based control methodology for power electronic converters," in *Proc. IEEE Green Technol. Conf. (GreenTech)*, Apr. 2021, pp. 81–88, doi: 10.1109/GreenTech48523.2021.00024.

[38] 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.

[39] Y. Wan, T. Dragicevic, N. Mijatovic, C. Li, and J. Rodriguez, "Reinforcement learning based weighting factor design of model predictive control for power electronic converters," in *Proc. IEEE Int. Conf. Predictive Control Electr. Drives Power Electron. (PRECEDE)*, Nov. 2021, pp. 738–743, doi: 10.1109/PRECEDE51386.2021.9680964.

[40] J. Wang, R. Yang, and Z. Yao, "Efficiency optimization design of three-level active neutral point clamped inverter based on deep reinforcement learning," in *Proc. IEEE 6th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Nov. 2022, pp. 605–610, doi: 10.1109/EI256261.2022.10117037.

[41] P. Qashqai, K. Al-Haddad, and R. Zgheib, "A new model-free space vector modulation technique for multilevel inverters based on deep reinforcement learning," in *Proc. 46th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2020, pp. 2407–2411.

[42] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133653–133667, 2019.

[43] D. Zhao, H. Wang, K. Shao, and Y. Zhu, "Deep reinforcement learning with experience replay based on SARSA," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2016, pp. 1–6, doi: 10.1109/SSCI.2016.7849837.

[44] Y. Zhang and T. Wang, "Applying value-based deep reinforcement learning on KPI time series anomaly detection," in *Proc. IEEE 15th Int. Conf. Cloud Comput. (CLOUD)*, Jul. 2022, pp. 197–202, doi: 10.1109/CLOUD55607.2022.00039.

[45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[46] R. Mukhopadhyay, S. Bandyopadhyay, A. Sutradhar, and P. Chattopadhyay, "Performance analysis of deep Q networks and advantage actor critic algorithms in designing reinforcement learning-based self-tuning PID controllers," in *Proc. IEEE Bombay Sect. Signature Conf. (IBSSC)*, Jul. 2019, pp. 1–6.

[47] Y.-T. Liu, J.-M. Yang, L. Chen, T. Guo, and Y. Jiang, "Overview of reinforcement learning based on value and policy," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Aug. 2020, pp. 598–603, doi: 10.1109/CCDC49329.2020.9164615.

[48] D. Dutta and S. R. Upreti, "A survey and comparative evaluation of actor-critic methods in process control," *Can. J. Chem. Eng.*, vol. 100, no. 9, pp. 2028–2056, Sep. 2022, doi: 10.1002/cjce.24508.

[49] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao, "Deep reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 28, 2022, doi: 10.1109/TNNLS.2022.3207346.

[50] *Reinforcement Learning Agents—MATLAB & Simulink*. Accessed: Mar. 15, 2023. [Online]. Available: https://www.mathworks.com/help/reinforcement-learning/ug/create-agents-for-reinforcement-learning.html

POURIA QASHQAI (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Isfahan, Isfahan, Iran, in 2013, and the M.Sc. degree in power electronics engineering from the Babol Noshirvani University of Technology, Babol, Iran, in 2015. He is currently pursuing the Ph.D. degree in power electrical engineering with École de technologie supérieure (ÉTS), University of Québec, Montreal, QC, Canada.

He is a member of Groupe de recherche en électronique de puissance et commande industrielle (GRÉPCI), Montreal. His research interests include the development of intelligent and robust control methods for multi-level power electronic converters, applications of artificial intelligence in power electronics, the development of real-time simulation models for electrical vehicles, and renewable energy applications using machine learning.

MOHAMMAD BABAIE (Graduate Student Member, IEEE) received the B.Sc. degree in electronic engineering from the Sepahan Institute of Science and Technology of Higher Education, Isfahan, Iran, in 2013, and the M.Sc. degree in control engineering from the Babol Noshirvani University of Technology, Babol, Iran, in 2016. He is currently pursuing the Ph.D. degree in power electrical engineering with École de technologie superieure, Montreal, QC, Canada.

He has authored or coauthored several journal and conference papers in the field of control and power electronics and holds five patents. His research interests include developing variable structure control theory, modeling, control of power electronics converters using robust, adaptive, intelligent control techniques, developing artificial neural network training strategies in power systems, and real-time control based on the FPGA and 32-bit MCUs for power electronic converters.

RAWAD ZGHEIB (Member, IEEE) received the Ph.D. degree in electrical engineering from École de technologie supérieure, Montreal, Canada, the master's degree in mobility and electric vehicles from École nationale supérieure d'Arts et Métiers, Paris, France, and the master's degree in research and electrical engineering from École supérieure d'ingénieurs de Beyrouth, Université Saint Joseph, Beirut, Lebanon.

He is currently a Research and Development Project Manager with Institut de recherche d'Hydro-Québec (Research Institute of Hydro-Québec), QC, Canada, in the fields of energy system resilience, energy transition, and virtualization tools. His research interests include the development of collaborative simulation tools that contributes to a global virtualization of the electrical networks, while considering the energy transition and digital transformation, modeling electrical network components in electrical, telecommunications, automation tools, and implementing case studies that examine the performance of power electronics and power system simulation tools and evaluate their impact on the resilience of the grid.

KAMAL AL-HADDAD (Life Fellow, IEEE) received the B.Sc.A. and M.Sc.A. degrees from Université du Quebec à Trois-Rivières, Trois-Rivières, QC, Canada, in 1982 and 1984, respectively, and the Ph.D. degree from the Institute National Polytechnique of Toulouse, Toulouse, France, in 1988.

Since June 1990, he has been a Professor with the Electrical Engineering Department, École de technologie supérieure, Montreal, QC, Canada, where he has been the Senior Canada Research Chair of the Electric Energy Conversion and Power Electronics, since 2002. He is currently a consultant and has established a very solid link with many Canadian and international industries working in the field of power electronics, electric transportation, aeronautics, and telecommunications. He successfully transferred and implemented dozens of technologies to Canadian and international companies. His research interests include highly efficient static power converters, harmonics and reactive power control using hybrid filters, voltage-level multiplier, resonant and multilevel converters, including the modeling, control, and development of prototypes for various industrial applications in electric traction, renewable energy, power supplies for drives, and telecommunication. He is a member of the Academy of Sciences, a fellow of the Royal Society of Canada, and a Fellow Member of the Canadian Academy of Engineering. He was a recipient of the 2014 IEEE IES Dr.-Ing. Eugene Mittelmann Achievement Award and the 2023 Medal in Power Engineering. He is the IEEE 2023-2024 Division VI Director. He is an Associate Editor of IEEE Transactions on Industrial Informatics and an IES Distinguished Lecturer. He was an IEEE IES President, from 2016 to 2017.