Delft University of Technology

Reinforcement Learning Based Control of Grid-Connected PUC5 Inverter

Kermansaravi, Azadeh; Alquennah, Alamera Nouran; Lekić, Aleksandra; Trabelsi, Mohamed; Ghrayeb, Ali; Abu-Rub, Haitham; Vahedi, Hani

DOI
10.1109/IECON55916.2024.10905177

Publication date
2024

Document Version
Final published version

Published in
IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society, Proceedings

Citation (APA)
Kermansaravi, A., Alquennah, A. N., Lekić, A., Trabelsi, M., Ghrayeb, A., Abu-Rub, H., & Vahedi, H. (2024). Reinforcement Learning Based Control of Grid-Connected PUC5 Inverter. In *IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society, Proceedings* (IECON Proceedings (Industrial Electronics Conference)). IEEE. https://doi.org/10.1109/IECON55916.2024.10905177

IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society | 978-1-6654-6454-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IECON55916.2024.10905177

Reinforcement Learning Based Control of Grid-Connected PUC5 Inverter

Azadeh Kermansaravi, *Member, IEEE*, Alamera Nouran Alquennah *Member, IEEE*,
Aleksandra Lekić, *Senior Member, IEEE*, Mohamed Trabelsi, *Senior Member, IEEE*,
Ali Ghrayeb, *fellow, IEEE*, Haitham Abu-rub *fellow, IEEE*, Hani Vahedi, *Senior Member, IEEE*

Delft University of Technology, Delft, Netherlands
Texas A&M University, College Station, TX, USA
Kuwait College of Science and Technology, Kuwait, Kuwait,
Texas A&M University at Qatar, Doha, Qatar

Z.Kermansaravi@tudelft.nl, alamera_nq@tamu.edu, A.Lekic@tudelft.nl, m.trabelsi@kcst.edu.kw,
ali.ghrayeb@qatar.tamu.edu, haitham.abu-rub@qatar.tamu.edu, H.Vahedi@tudelft.nl

***Abstract*—In this paper, a Reinforcement Learning controller (RLC) is designed and implemented on a 5-level Packed U-Cell (PUC5) grid-connected inverter to control the injected current flowing into the electric network.The RL agent is trained using a Proportional-Integral (PI) reward function to optimize its control strategy. Moreover, the voltage balancing of the auxiliary capacitor in PUC5 is separated from the RL controller and integrated into the switching algorithm to reduce the training burden. This modification reduces the observation inputs required for RL training, significantly shorten the training time. Simulation studies conducted in Matlab/Simulink evaluate the performance of the proposed RL controller, demonstrating robust dynamic response and accurate tracking of reference signals across different operational conditions.**

***Index Terms*—Reinforcement Learning, PUC5, Grid-Connected Inverter, AI controller**

I. INTRODUCTION

To achieve the clean energy transition, polluting applications using fossil fuels should be electrified to be powered by green energy resources. Power Electronics Converters play a key role in this aspect where they are currently used in various applications such as PV, Wind, battery storage and charging, electric vehicles, hydrogen production (electrolyzer), fuel cells, etc [1], [2]. A significant increase in installing power converters necessitates the development of devices with better performance and efficiency. Multilevel Inverters are growing gradually and finding their way into the industry due to their inherent low harmonic pollution, which results in reduced passive filter size and increased efficiency. Cascaded H-Bridge (CHB), Modular Multilevel Converter (MMC), and Neutral Point Clamped (NPC) are among the topologies used for high-power medium-voltage applications [3]. The 5-level packed-U-cell (PUC5) is a recent invention that has attracted attention from the industry [4]. Currently, it is commercialized for EV charging by the dcbel company [5]. As shown in Fig. 1, compared to a full bridge inverter, it has two more switches and one auxiliary capacitor, which are necessary to generate a quasi-sine wave 5-level output voltage with low harmonic distortion. It has been demonstrated in both stand-alone and grid-connected modes of operation [6], [7]. Similar to the existing converters in the market, it uses only a single DC source, making it suitable for applying all types of controllers [8]–[13]. Model Predictive Control (MPC) and Sliding Mode Control (SMC) have previously been designed and applied to the PUC5 inverter to inject a desired amount of current/power with a regulated power factor into the grid [14], [15]. Although those model-based current controllers could achieve higher accuracy, they always depend on the system parameters and cannot cover a wide range of variations. On the other hand, applying a model-independent PI controller on a single-phase inverter usually leads to a steady-state error in the current amplitude. A comparison of different control techniques for MLIs is presented in [16]–[18], indicating the current trend of

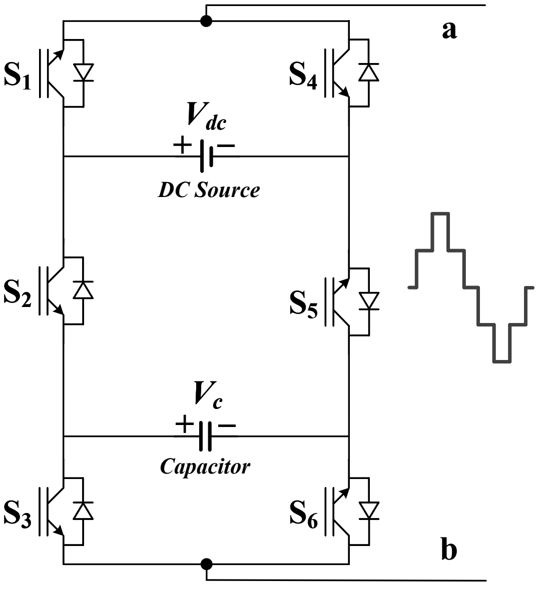

Fig. 1. PUC5 Inverter topology

AI-based solutions.

Reinforcement learning control (RLC) is a model-free control algorithm that can learn the proper control policy and achieve the desired control objectives in a multilevel inverter system [18]. Recent studies have explored the application of reinforcement learning to grid-connected inverters. [19] presented a modern RL-based discrete-time control methodology for power electronic converters, showcasing the advantages of RL over classical PID control methods. In [20] the authors designed a deep reinforcement learning (DRL) controller for a grid-connected inverter with an LCL filter, demonstrating improved performance and robustness compared to traditional control methods. This highlights the potential of DRL in handling nonlinearities and uncertainties in inverter systems. Similarly, in [21] the authors developed a reinforcement learning-based controller specifically for three-phase grid-connected inverters, showcasing enhanced performance and adaptability. In addition, [22] explores the use of a deep deterministic policy gradient (DDPG) algorithm for controlling a three-phase inverter, further underlining the versatility and effectiveness of RL approaches in power electronics applications. These studies underscore the applicability of RL in developing advanced control strategies for power electronic converters. These studies underscore the applicability of RL in developing advanced control strategies for power electronic converters.

Furthermore, recent studies have demonstrated RLC's application for PUC5 and PUC7 using various RLC algorithms [23] and [24]. These studies have shown that RLC is a competitive control algorithm for such systems. It learns solely through the agent's interaction with the system without the need for collecting data for a specific operating scenario or depending on the model. Furthermore, wide operating scenarios can be included in the designed environment of the RLC, exposing the agent during the training phase to various randomness in the system.

However, this comes with the price of a high computation burden and a lengthy training phase. Additionally, proper design of the RLC components is required, including the training environment with all the random scenarios, the informative reward function that will guide the agent to learn the control objectives, and a deep neural network (NN) with all the hyper-parameters. Despite all these learning requirements, in the implementation and testing phase, the trained agent can be applied as conventional neural network models without the RLC elements used in training. Therefore, this paper proposes a modified RLC for a single-phase grid-connected PUC5 that accelerates the training effort, compared to the previously proposed RLC, by reducing the action space and decoupling the control objectives of the PUC5 system by concentrating the agent training on tracking the reference current signal only. This is done by introducing a PI-based reward function in the RL controller design. The capacitor voltage regulating objective is achieved through the redundant switching states.

PUC5 topology and its auxiliary capacitor voltage balancing through the switching algorithm are briefly explained in section II. The RL controller design procedure is described in section III, which elaborates on the PI-based reward function. Simulations in Matlab/Simulink are done to investigate the steady-state and dynamic performance of the proposed controller, and results are shown and discussed in section IV.

II. PUC5 TOPOLOGY, SWITCHING AND VOLTAGE BALANCING

PUC5 inverter was invented by Vahedi et al. in 2014 [4]. It consists of 6 switches and 2 DC links. By controlling the capacitor voltage $V_c = E$ at half of the DC source level $V_{dc} = 2E$, a 5-level voltage waveform can be generated at the output. Having 3 pairs of complementary switches, 8 switching states will be derived, as shown in Table I.

TABLE I
SWITCHING STATES OF PUC5 INVERTER

Switching State	S_1	S_2	S_3	V_{out}	Capacitor Charge/Discharge
1	1	0	0	$+2E$	–
2	1	0	1	$+E$	Charge
3	1	1	0	$+E$	Discharge
4	1	1	1	0	–
5	0	0	0	0	–
6	0	0	1	$-E$	Discharge
7	0	1	0	$-E$	Charge
8	0	1	1	$-2E$	–

The redundant states (2-3 and 6-7) are utilized to balance the capacitor voltage. They generate the same voltage level at the output while connecting different components to make a path for current. For instance, state 2 connects the DC source and capacitor in series, which charges the capacitor. Also, state 3 directly connects the capacitor to the output, discharging it. Therefore, the capacitor voltage can be kept at the desired level by selecting the appropriate redundant state. This can be implemented by modifying the switching algorithm to use feedback from the capacitor voltage. Thus, if voltage level +E or -E is supposed to be generated at the output, the voltage feedback will be used to select between charging or discharging states. This will ensure the capacitor voltage is regulated at the desired level without employing an additional voltage controller.

III. RL-BASED CONTROLLER DESIGN

As depicted in Fig. 2, the AC current i_{ac} should be controlled with a desired amplitude and phase in synchronization with grid voltage v_{ac}. The RL agent is designed to dynamically select from a set of switching configurations to optimize the performance of the PUC5 inverter when injecting the current into the grid. The switching configurations are considered as the action space for the RL agent.

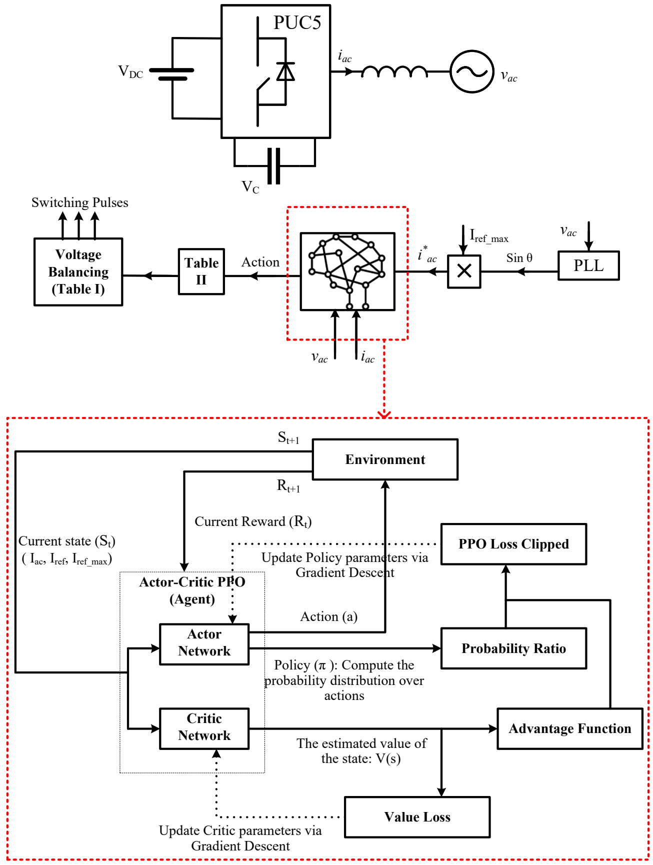

Fig. 2. Grid-Connected PUC5 inverter with RL Controller

A. Observation and Action Space

The observation space, representing the normalized state information of the system, and the action space, representing the switching configurations, are defined as follows:

1) Observation Space: The observation space consists of three normalized parameters: i_{ac}, i^*, and I^*_{max}, where i_{ac} is the AC current, i^* is the reference current as a unit sine wave synchronized with v_{ac}, and I^*_{max} is the maximum reference current. Hence, the observation space is represented as a continuous space with a dimension of 3×1. These parameters are normalized between -1 and 1 to ensure stability and consistency in the learning process. The normalization is implemented in MATLAB rlNumericSpec function definition, which specifies the lower and upper limits for the observation space.

2) Action Space: As mentioned in the abstract, one contribution in this paper is to reduce the action space and training time by decoupling the capacitor voltage control from the current control. Therefore, the capacitor voltage will be balanced through redundant states, as Section II describes. Therefore, a modified and simpler switching table can be designed for the action space. The action space is discrete and consists of five switching configurations. These configurations are labeled as follows:

TABLE II
ACTION SPACE OF RL CONTROL FOR PUC5 INVERTER

Action	V_{out}
A1	$+2E$
A2	$+E$
A3	0
A4	$-E$
A5	$-2E$

It should be noted that:

- Action A2 corresponds to switching states 2&3.
- Action A3 corresponds to switching states 4&5.
- Action A4 corresponds to switching states 6&7.

These selections will be performed in a switching algorithm as an integrated voltage balancing unit.

B. State Space

The state space for the RL agent includes the normalized observation parameters described previously. These normalized parameters provide a consistent basis for the RL agent to make decisions, ensuring that the learning process remains stable and efficient. Specifically, the state space includes the AC current (i_{ac}), the reference current (i^*), and the maximum reference current (I^*_{max}).

C. Reward Function

The objective of the reinforcement learning-based control design is to minimize the error between the reference current i^* and the actual current i_{ac} while penalizing overshooting and controlling the input effort. The reward function incorporates several components to ensure effective control:

- Proportional Error: The proportional error e_p between the reference current i^* and the actual current i_{ac} is calculated as follows:

$$e_p = i^* - i_{ac} \tag{1}$$

  This term encourages the RL agent to minimize the difference between the desired and actual current values.
- Normalization of Errors: The errors are normalized to account for the varying magnitudes of i^*:

$$\text{normalizedError} = \begin{cases} 0 & \text{if } i^* = 0 \\ \frac{e_p}{i^*} & \text{otherwise} \end{cases} \tag{2}$$

$$\text{normalizedIntegralError} = \begin{cases} 0 & \text{if } i^* = 0 \\ \frac{integralError}{i^*} & \text{otherwise} \end{cases} \tag{3}$$

  Normalizing the errors helps maintain consistency in the reward function regardless of the reference current's magnitude, which is crucial for stable learning.
- Penalties for Overshooting: An overshoot penalty is introduced to penalize instances where the actual current exceeds

the reference current, which is undesirable in control systems:

$$\text{overshootPenalty} = \begin{cases} |i_{ac} - i^*| & \text{if } |i_{ac}| > |i^*| \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

This term helps reduce the overshoot, which can lead to instability and inefficiency in the system.

- Control Input: The control input u represents the control action, and its penalizing helped reduce the training time and the agent's convergence. However, further investigation on the effect of this penalty term will be conducted in future work.
- Reward Function: The final reward function is a combination of the absolute values of the proportional and integral errors, the overshoot penalty, and the control input effort:

$$\begin{aligned} r_k = &-|e_p| - \alpha|integralError| \\ &- \text{overshootPenalty} - \beta u + 0.01 \end{aligned} \quad (5)$$

Where α is a weighting coefficient for the integral term, β is a weighting factor for the control action u, and a small positive reward is added to encourage exploration. It is worth mentioning that the optimal weighting coefficient and the exact impact of each reward term are part of future work.

The design of this reward function ensures that the RL agent aims to minimize the proportional error e_p, account for the accumulated error through the integral term, and penalize the overshooting to maintain system stability.

By maximizing the cumulative reward, the RL agent effectively learns to improve the tracking of the reference current, minimize oscillations, and maintain overall system stability and efficiency.

D. RL Architecture and Learning Algorithm

The control design employs a reinforcement learning (RL) framework using the actor-critic architecture to achieve efficient control for the single-phase grid-connected PUC5 inverter [25].

The Proximal Policy Optimization (PPO) algorithm is selected for its suitability for discrete control tasks. The PPO learning diagram outlines the key components and steps involved in training an agent through reinforcement learning. The algorithm parameters are configured with an experience horizon of 1000 steps, a mini-batch size of 512, and a clip factor of 0.2 to stabilize training. Each training epoch incorporates three iterations to refine the actor and critic networks further. A discount factor of 0.99 ensures appropriate weighting of future rewards in decision-making.

The environment interacts with the agent by providing inputs such as the current AC current i_{ac}, reference current i^*, and maximum reference current I^*_{max}. It outputs observations, rewards, and the next state. The actor network takes the current state as input and produces a probability distribution over possible actions, governed by the policy $\pi_\theta(a \mid s)$. Simultaneously, the critic network estimates the state value $V_\phi(s)$ to evaluate the expected return. During action selection, an action is sampled from the actor network's distribution and executed in the environment. The environment then provides the next state and a reward signal based on the action taken. Experiences comprising tuples of state, action, reward, and next state are stored in a replay buffer for subsequent training. Given that r is the reward function, V is the value function, S is the current state, $S^{'}$ is the next state, a is the taken action and γ is the discount factor, the advantage function $A(s,a) = r(s,a) + \gamma V(s') - V(s)$ can be computed guiding the update process. The policy is updated using the PPO objective function, which clips the probability ratio to ensure stability, while the critic network is updated to minimize the value loss [26]. This iterative process of observing states, predicting actions, executing them, evaluating outcomes, and updating policies is repeated over multiple episodes until convergence is achieved.

The RL agent is trained within a Simulink environment tailored to simulate the behavior of the PUC5 inverter under varying conditions. The training process spans up to 10,000 episodes, with each episode capped at one complete sine wave (333 steps) to balance training time and computational resources. Training progress is monitored and evaluated through training progress plots, providing transparency and insights into the agent's performance [27].

By integrating this RL architecture and learning algorithm, the control strategy for the grid-connected PUC5 inverter aims to optimize operational efficiency while leveraging advanced computational capabilities to achieve robust and adaptive control in dynamic grid environments.

IV. SIMULATION RESULTS AND DISCUSSION

The grid-connected PUC5 inverter (shown in Fig. 2) has been simulated in Matlab/Simulink. Simulation parameters are listed in Table III. The proposed RL-based controller has been applied, and the results have been captured for different scenarios. The AC current is multiplied by 10 to show it along with AC voltage waveform in the same window.

TABLE III
SIMULATION PARAMETERS

Grid Voltage (RMS)	$120V$
DC Voltage (V_{dc})	$200V$
Inductor (L)	$5mH$
Auxiliary Capacitor	$200\mu F$
Sampling Time	$50\mu s$

The steady-state results are illustrated in Fig. 3 where the current amplitude is set to 10A and synchronized with the grid voltage. V_C is balanced at 100V as half of the DC source amplitude. The THD% of the i_{ac} is measured at 2.5%. Moreover, the 5-level voltage waveform is generated smoothly, which shows that the proper actions are selected by the RL controller.

To investigate the dynamic performance of the designed controller, various changes are intentionally applied to system

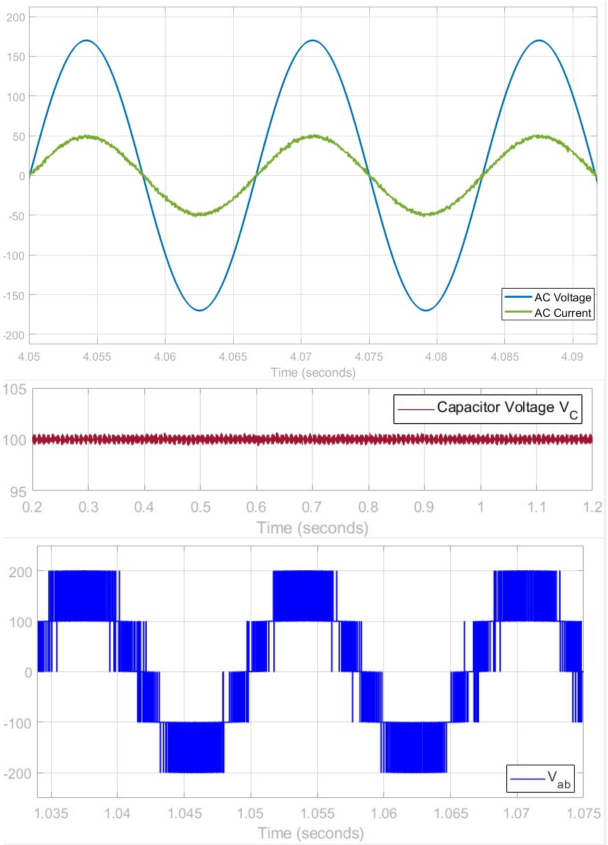

Fig. 3. steady state results pf PUC5 grid-connected inverter with RL-based current controller

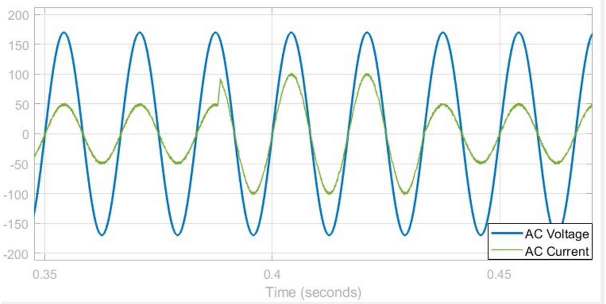

Fig. 4. Results during a change in current amplitude from 5A to 10A

parameters. First, the current amplitude is changed to 5A, and the results are depicted in Fig. 4, proving the controller's precise tracking of the reference value.

Second, the power factor is changed from 1 to 0.86 to exchange reactive power with the grid. Results are illustrated in Fig. 5 with an acceptable control on the current. This highlights the capability of the designed controller, not only to feed the grid with a unity power factor, but to compensate for the grid changes.

In another test, the AC voltage varies from 170V to 140V peak, and the results are shown in Fig. 6. Those waveforms confirm that the implemented controller could precisely perform the job even during such a disturbance.

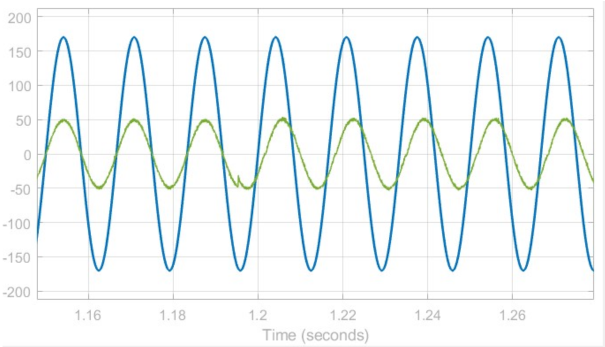

Fig. 5. Results during a change in power factor from 1 to 0.86

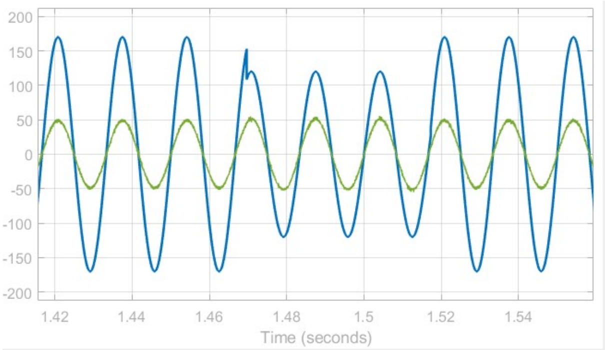

Fig. 6. Results during a change in AC voltage variation from 120V RMS to 100V RMS

V. CONCLUSION

In this paper, a model-free RL controller has been designed and implemented on a single-phase grid-connected PUC5 inverter to regulate the injected current using a PI-based reward function. This novel design concentrates the control effort on reducing the current error in the grid-tied system. This reduces the learning time as the voltage balancing of the PUC5 inverter was done separately in the switching algorithm leading to a reduced action space and RL training effort. Simulation results confirmed the acceptable performance of the proposed controller during steady state and transients of the system, such as changes in current amplitude and AC and DC voltage levels. This work showed that the RL is a potential candidate as a model-independent controller to be applied to power electronics converters, especially the multilevel topologies.

VI. ACKNOWLEDGEMENT

The project was funded by Kuwait Foundation for the Advancement of Sciences (KFAS) under project code: CN23-13EE-1882.

REFERENCES

[1] H. Abu-Rub, M. Malinowski, and K. Al-Haddad, *Power electronics for renewable energy systems, transportation and industrial applications*. John Wiley & Sons, 2014.

[2] N. Jalilian, G. D. Najafpour, and M. Khajouei, "Macro and micro algae in pollution control and biofuel production – a review," *ChemBioEng Reviews*, vol. 7, no. 1, pp. 18–33, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cben.201900014

[3] M. Trabelsi, H. Vahedi, and H. Abu-Rub, “Review on single-dc-source multilevel inverters: Topologies, challenges, industrial applications, and recommendations,” *IEEE Open Journal of the Industrial Electronics Society*, vol. 2, pp. 112–127, 2021.

[4] H. Vahedi and K. Al-Haddad, “Method and system for operating a multilevel electric power inverter,” Mar. 20 2018, uS Patent 9,923,484.

[5] H. Vahedi, M.-a. Forget, and P. Ibrahim, “Electric vehicle battery charger,” Jan. 11 2024, uS Patent App. 18/473,553.

[6] H. Vahedi, P.-A. Labbé, and K. Al-Haddad, “Sensor-less five-level packed u-cell (puc5) inverter operating in stand-alone and grid-connected modes,” *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 361–370, 2015.

[7] H. Vahedi, M. Sharifzadeh, and K. Al-Haddad, “Topology and control analysis of single-dc-source five-level packed u-cell inverter (puc5),” in *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2017, pp. 8691–8696.

[8] K. Rayane, M. Bougrine, A. Krama, A. Benalia, M. Trabelsi, and H. Abu-Rub, “Average model-based feedforward and feedback control for puc5 inverter,” *IEEE Access*, vol. 8, pp. 172 962–172 971, 2020.

[9] H. Makhamreh, M. Trabelsi, O. Kükrer, and H. Abu-Rub, “A lyapunov-based model predictive control design with reduced sensors for a puc7 rectifier,” *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1139–1147, 2020.

[10] A. Bhanuchandar and B. K. Murthy, “A new floor function single-carrier-based modulated model predictive current control technique for single-phase puc5 inverter topology,” *International Journal of Circuit Theory and Applications*, vol. 52, no. 2, pp. 812–834, 2024.

[11] M. Babaie, M. Sharifzadeh, H. Y. Kanaan, and K. Al-Haddad, “Switching-based optimized sliding-mode control for capacitor self-voltage balancing operation of seven-level puc inverter,” *IEEE Transactions on Industrial Electronics*, vol. 68, no. 4, pp. 3044–3057, 2020.

[12] H. Vahedi, M. Trabelsi, H. Vahedi, and M. Trabelsi, *Single-dc-source multilevel inverters*. Springer, 2019.

[13] I. Harbi, J. Rodriguez, A. Poorfakhraei, H. Vahedi, M. Guse, M. Trabelsi, M. Abdelrahem, M. Ahmed, M. Fahad, C.-H. Lin *et al.*, “Common dc-link multilevel converters: topologies, control and industrial applications,” *IEEE Open Journal of Power Electronics*, 2023.

[14] M. Trabelsi, S. Bayhan, K. A. Ghazi, H. Abu-Rub, and L. Ben-Brahim, “Finite-control-set model predictive control for grid-connected packed-u-cells multilevel inverter,” *IEEE Transactions on Industrial Electronics*, vol. 63, no. 11, pp. 7286–7295, 2016.

[15] K. Rayane, M. Bougrine, H. Abu-Rub, A. Benalia, and M. Trabelsi, “Average model-based sliding mode control for quality improvement of a grid-connected puc inverter,” *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 11, no. 4, pp. 3765–3774, 2023.

[16] I. Harbi, J. Rodriguez, E. Liegmann, H. Makhamreh, M. L. Heldwein, M. Novak, M. Rossi, M. Abdelrahem, M. Trabelsi, M. Ahmed *et al.*, “Model-predictive control of multilevel inverters: challenges, recent advances, and trends,” *IEEE Transactions on Power Electronics*, vol. 38, no. 9, pp. 10 845–10 868, 2023.

[17] M. Trabelsi, A. N. Alquennah, and H. Vahedi, “Review on single-dc-source multilevel inverters: Voltage balancing and control techniques,” *IEEE Open Journal of the Industrial Electronics Society*, vol. 3, pp. 711–732, 2022.

[18] Y. Fujimoto, A. Kaneko, Y. Iino, H. Ishii, and Y. Hayashi, “Challenges in smartizing operational management of functionally-smart inverters for distributed energy resources: A review on machine learning aspects,” *Energies*, vol. 16, no. 3, p. 1330, 2023.

[19] D. Alfred, D. Czarkowski, and J. Teng, “Reinforcement learning-based control of a power electronic converter,” *Mathematics*, vol. 12, no. 5, p. 671, 2024.

[20] A. Rajamallaiah, S. P. K. Karri, M. L. Alghaythi, and M. S. Alshammari, “Deep reinforcement learning based control of a grid connected inverter with lcl-filter for renewable solar applications,” *IEEE Access*, 2024.

[21] R. kumar Mahto, P. Mishra, and A. Alam, “Design of controller for three-phase grid-connected inverter using reinforcement learning,” in *2024 1st International Conference on Innovative Sustainable Technologies for Energy, Mechatronics, and Smart Systems (ISTEMS)*. IEEE, 2024, pp. 1–6.

[22] C. Xiang, X. Zhang, T. Qie, T. K. Chau, J. Ye, Y. Yu, H. H. C. Iu, and T. Fernando, “A novel deep deterministic policy gradient assisted learning-based control algorithm for three-phase dc/ac inverter with an rl load,” *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 11, no. 6, pp. 5529–5539, 2022.

[23] A. N. Alquennah, M. Chida, T. Zamzam, and M. Trabelsi, “Reinforcement learning based controller for grid-connected puc pv inverter,” in *IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society*, 2023, pp. 1–6.

[24] A. N. Alquennah, M. Chida, T. Zamzam, M. Trabelsi, A. Ghrayeb, and S. Khatri, “Double deep q networks reinforcement learning based controller for grid-connected 7-level puc inverter,” in *2024 4th International Conference on Smart Grid and Renewable Energy (SGRE)*, 2024, pp. 1–6.

[25] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, “A survey of actor-critic reinforcement learning: Standard and natural policy gradients,” *IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[26] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Advances in neural information processing systems*, vol. 12, 1999.

[27] Y. Li, “Deep reinforcement learning: An overview,” *arXiv preprint arXiv:1701.07274*, 2017.