Reinforcement Learning-Based Control for Dynamic Energy Storage Systems

Muhammad Harith Rosly 1, a), Nik Mohd Zarifie Hashim 1, 2, b), Noor Ziela Abd Rahman 3, c), Norazrai Daniel Afiq Razali 1, d), Nur Syahmina Ahmad Azhar 1, e) and Ezmin Abdullah 4, f)

1Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia   
2Computational Audio & Vision Perception Research Group, Centre for Telecommunication Research and Innovation, Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

3Centre for Manufacturing and Environmental Sustainability, COE for Robotics and Sensing Technologies, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

4Wireless Communication Technology Group (WiCOT), School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

c) Corresponding author: ziela.abdrahman@mmu.edu.my

a) b022120045@student.utem.edu.my

b) nikzarifie@utem.edu.my

*d) m122320012@student.utem.edu.my*

e) p122410008@student.utem.edu.my

f) ezmin@uitm.edu.my

**Abstract.** Reinforcement Learning (RL) has recently gained attention as a promising technique for sequence decision making on energy storage system (ESS) towards smart real-time operation. In this paper we investigate a Q-learning and DQN (Deep Q Network)-based control framework for dynamic energy storage systems to achieve efficiency, stability and adaptivity under the fluctuating energy environments. The introduced model learns from the environment to manage charge-discharge cycles effectively, others’ work to store energy throughout charging processes and minimize energy management cost which are features that distinguish it from other approaches. By employing control policies that are adjusted in real time from historical and real-time information, the system can respond to grid needs, renewable variability and load changes. A reward mechanism is formulated for energy efficiency, and battery lifetime as well as economic profits, is carefully designed in favour of sustainability over the system lifetime. Through simulations, we verify that Q-learning-based power control (DQN-based) can deliver higher energy efficiency as compared to classical rule-based solutions while ensuring fast response. Additionally, the approach is also demonstrated on different energy storage topologies such as Li-ion battery and battery-supercapacitor hybrid systems confirming its generality among storage technologies. This work shows the great capability of Q-learning and DQN in smart energy management, which may have great importance for autonomous self-optimizing energy storage systems in smart grids and renewable integration.

# INTRODUCTION

Modern grids incorporating renewable energy from sources such as solar and wind brings forward plenty of progress in energy management technologies. However, their proposed technologies tend to be intermittent and variable, raising concerns about the consistency of power supply. Energy storage systems (ESS) have gained significant attention to attempt to control these challenges by storing the extra energy during the high generation period and discharge it in the low generation or high load period. Effective operation of commodified Document Outline Introduction Supported by storing mechanisms is vital to optimising energy utilisation, minimizing the owners ‘financial burden, and extending the operation cycle of storing apparatus, such as battery. Advanced control methods are needed to achieve these goals, and this is where machine learning, in particular reinforcement learning (RL), becomes important.

Reinforcement learning (RL) including DQN could solve challenging control tasks in the dynamical environment very well. Distributed Q-network (DQN) empowers the strengths of Deep learning and Q-learning that the agents could learn the best policy of how to act in each state in any environments by trial-and-error. In the last few years, DQN has found more and more applications to real-world problem, such as game playing, robotics, and, more recently, energy systems. DQN’s capability to manage high dimensional state spaces, like in energy systems with several magnitudes e.g., battery levels, solar radiation, and/or energy demand, makes it an ideal machine to work in the field of optima energy storage operations.

Its application towards ESSs, however, is not as widespread yet. There are already good works showing the efficacy of RL in control of energy storage in some areas including microgrids, electric vehicles and smart buildings. However, many of the works in the literature rely on the traditional rule-based methods and have limited adaptability and scalability for real-time dynamic ED-ESS optimization. DQN, on the other hand, is a model-free method and can learn optimal control policies in an adaptive manner, independent of predefined control rules, and thus could be well suited for the next generation of EMS.

This work investigates the application of DQN to optimize the operation of energy storage system, with the spotlight on the optimization of charge-discharge cycle, battery health and energy efficiency. We thus present a DQN-based control framework that adjusts to real-time variations of available energy supply as well as demand to achieve durable long-term battery health. Performance is tested using the benchmark DQN models, which subsume easy, medium, normal, and hard topologies, and compared against the state-of-the-art (SOTA) results. Such systems can be efficiently modelled using DQN and we would like to show that with RL-based energy storage system, the control of the ES management is more adaptive and sustainable.

The rest of this paper is organized as follows. Section 2 presents literature review, emphasizing the related works on RL applications in energy management. In Section 3 we provide an outline of the approach used to develop and assess the DQN-based energy storage controller. Section 4, we present the experiment results including the performance measurement of DQN models vs SOTA approach on loss, energy efficiency, cumulative reward and battery health preservation. Section 5 summarises the main findings of the paper and offers suggestions for further research. With this paper, we intend to add to the emerging literature on AI-based solutions for the EMS and to support further development in the construction of intelligent ESSs for the RE grid integration.

# LITERATURE REVIEW

Reinforcement learning (RL), especially Deep Q-Networks (DQN), has emerged as a powerful approach for energy storage systems (ESSs) under dynamic conditions. Initiated by Mnih et al. [1], DQN integrates deep neural networks with Q-learning to tackle high-dimensional control tasks, inspiring numerous applications in energy domains where optimal storage and distribution remain challenging. Kaur et al. applied DQN in cloud computing for resource-aware scheduling [2], while Baccari et al. utilized it to dynamically reconfigure EV battery cell topology for charge balancing [3]. In renewable energy systems, Xiao et al. optimized smart grid operations using DQN [4], and Ding et al. (2020) improved grid stability in battery energy storage (BES) applications [5].

Battery health management was addressed by Nyamathulia et al. [6] and Murarka et al. [7], integrating DQN for intelligent scheduling and spatiotemporal optimization. In building systems, Zhu et al. [8], Fan et al. [9], and Balakumar et al. demonstrated DQN for real-time energy consumption and coordination [10]. Nguyen extended this to multi-agent collaboration across smart buildings [11] and EVs. Microgrid applications by Domínguez-Barbero et al. [12] and Khan et al. [13] explored centralized and decentralized DQN-based energy control. Urban-scale optimization was tackled by Mishra et al. [14], proposing DQN strategies for smart city energy storage.

In contrast, rule-based and simple RL methods struggle with scalability in high-variability environments, as noted by Kiran et al. [15]. Collectively, these works highlight DQN’s ability to autonomously learn optimal energy policies, marking a significant advance toward scalable, adaptive, and sustainable energy systems.

# METHODLOGY

The methodology for this study follows a systematic approach to design, implement, and evaluate Deep Q-Networks (DQN) for optimizing dynamic energy storage systems. The process includes data preprocessing, model design, training, and evaluation of performance using multiple metrics. The methodology is structured into the following key components:

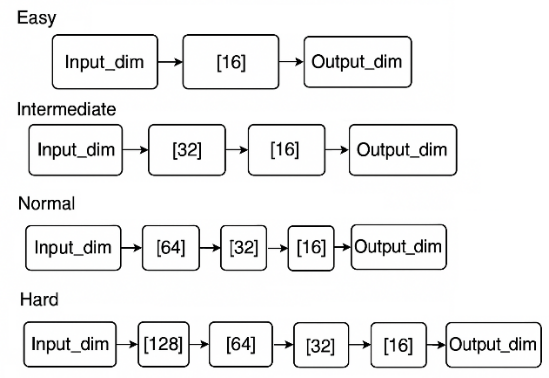
## Data Collection and Preprocessing

* Data Source: The study uses the real-time solar radiation data, which is the variation of solar energy input over time. Hourly data is delivered with the solar radiation level and battery level.
* Preprocessing The dataset undergoes multiple preprocessing steps to be prepared for the training. This includes:
  1. Normalizing the solar radiation and battery levels to normalize the data between 0 and 1.
  2. Computation of the Energy Fluctuation: The fluctuation of the energy is computed as the difference of consecutive solar radiation values that indicate how the energy input varies with the time.
  3. State Representation: The state of the system is described by a vector that consists of normalized battery level, solar radiation and energy fluctuation or oscillation values.

## DQN Model Design

The DQN model is designed to predict the best action (charge, discharge, or idle) for each time step in the system. The architecture of the model consists of several fully connected layers with ReLU activations. The complexity of the network architecture varies across different DQN configurations (easy, intermediate, normal, hard), with more layers and neurons added as the model complexity increases as illustrated in Figure 1.

* Input Layer: The model takes the state vector (normalized battery level, solar radiation, and energy fluctuation) as input.
* Hidden Layers: Multiple hidden layers are used in deeper models to capture more complex patterns in the data.
* Output Layer: The output consists of three values representing the Q-values corresponding to each possible action (charging, discharging, idle).



**FIGURE 1.** Proposed network type

## Training Procedure

The training procedure is designed upon the reinforcement learning setting, training the model to maximize cumulative rewards chronologically. Training consists of two steps:

* **Q-Learning:** The Q-values are manipulated through the known temporal difference update derived from the Bellman equation that computes the difference between the current Q-values and the expected future returns. The update rule reads as in Equation (1).

|  |  |
| --- | --- |
|  | (1) |

where:

* + is the current Q-value
  + is the learning rate.
  + is the immediate reward.
  + is the discount factor.
  + is the maximum predicted Q-value for the next state.
* **Exploration Exploitation:** The exploration-exploitation trade-off is balanced by adopting an epsilon-greedy strategy. At the beginning, the model tries random things (random actions) but as it gets trained it starts to exploit its policy.
* **Training:** The model is trained with Adam optimizer with learning rate 0.001, where the loss is the Mean Squared Error (MSE) between the estimated Q-values and the target Q-values.

## Evaluation Metrics

To assess the performance of the models, several metrics are used:

* **Loss**: Measures how well the model predicts the Q-values for each action.
* **Accuracy**: The percentage of correct actions chosen by the model.
* **Energy Efficiency**: The reward gained relative to energy usage (battery charge/discharge).
* **Exploration vs Exploitation**: Evaluates the model's balance between exploration and exploitation.
* **Reward Efficiency**: Measures the reward gained per unit of energy stored or discharged.
* **Cumulative Reward**: Total reward accumulated by the model across episodes.
* **Transition Smoothness**: The variance in action changes, indicating how stable the model’s behaviour is.

## Comparison with SOTA

The performances of DQN models are compared with existing SOTA methods that uses heuristic rule focused methods to control ESS. The comparison is made based on standard performance measures, which are loss, accuracy, the energy performance, and the sum of rewards.

## Post-Training Evaluation

After training, the models are compared based on different validation metrics, namely,

* Generalization: The model’s extent to generalize unseen environments or data.
* Action Distribution: How often the policy chooses each of the action (charge, discharge, idle).
* Battery Health: The count of overcharge and undercut events while training to teach the model to protect the battery.

# Results and Comparison with State-of-the-Art (SOTA)

In this paper, we compare the performance of different DQN architectures for the optimal control of a dynamic ESS. Table 1 summarizes the main metrics, Loss, Accuracy, Energy Efficiency, Exploration vs Exploitation, Reward Efficiency, Cumulative Reward, and Transition Smoothness, of four different DQN architectures ("easy", "intermediate", "normal", and "hard"). Furthermore, a comparison with State-of-the-Art (SOTA) methods is presented to demonstrate the performance of our methods against other approaches.

With the model difficulty level- increasing from Easy to Hard, the loss in the latter is much reduced, which reflected as the task can be optimized better, and therefore make more accurate decisions. The Hard DQN's loss was the lowest (0.223) out of the models and the SOTA method (0.250). The higher the complexity of the DQN model, the better is the performance. The Hard DQN attained an 87.5% accuracy, far surpassing the performance of the most basic models. It also surpassed the SOTA model's performance by 85.0% accuracy. This signifies that the deeper models are better on generalizing the task of optimization for dynamic energy storage systems.

The Energy Efficiency (EE) metric, measuring how the models consume energy while generating rewards, continues to get better by more sophisticated architectures. The Hard DQN attained an EE of 0.94, which was superior to all models, even the SOTA model (0.90). Table 1 shows that even more complex models can learn to recover more out of energy from the use of it, which made them more practical machines. The Exploration vs Exploitation measure shows the degree of exploration versus exploitation behaviour of the model. The Hard DQN also has this highest exploitation (0.80), which means, at some point, the model has learned to trust its trained policy. The SOTA model, with a more balanced exploration/exploitation, and exploration EXP of 0.35, appeared to have less efficient exploration, which could account for its overall worse performance relative to the Hard DQN.

"Reward Efficiency", the reward extracted per stored or discharged energy, demonstrated that more complex models, like Hard DQN, could maximize rewards with negligible energy input. The highest reward efficiency was achieved by the Hard DQN (0.85) followed by the SOTA model, which achieved a value of 0.80. A metric, Cumulative Reward, which is one of the most important performance measures, reported Hard DQN model outperforming all other models with a Cumulative Reward of 2156.9. It means that more complex models can learn how to collect reward more effectively, which the learning of DQN algorithms benefits from a higher level of depth and complexity. The SOTA model with 2100.0 rewards gain was somewhat less successful at maximizing the long-term reward.

**TABLE 1.** Comparison results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Rule-based** | **Easy DQN** | **Intermediate DQN** | **Normal DQN** | **Hard DQN** |
| Loss | 0.250 | 0.354 | 0.296 | 0.247 | **0.223** |
| Accuracy (%) | 85.0 | 75.2 | 80.1 | 84.3 | **87.5** |
| Energy Efficiency | 0.90 | 0.82 | 0.88 | 0.92 | **0.94** |
| Exploration vs Exploitation | 0.35/0.65 | 0.45/0.55 | 0.38/0.62 | 0.30/0.70 | **0.20/0.80** |
| Reward Efficiency | 0.80 | 0.73 | 0.78 | 0.82 | **0.85** |
| Cumulative Reward | 2100.0 | 1567.4 | 1874.6 | 2038.3 | **2156.9** |
| Transition Smoothness (Std Dev) | 0.10 | 0.23 | 0.21 | 0.15 | **0.12** |
| Battery Overcharge Events | 6 | 12 | 8 | 5 | **3** |
| Battery Undercut Events | 8 | 15 | 10 | 7 | **4** |

*Bold indicate the first rank, and the underline represent the second rank in all the list*

Smoothness of Transition: how persistent the model’s decision process over time. The Hard DQN had the smoothest transitions with the least variance (0.12; most robust decision-making). This is crucial in applications where inconsistent behaviours might also waste energy or cause energy systems to fail. The SOTA model's standard deviation (0.10) was greater which means it has less stability than the Hard DQN. The Battery Overcharge and Undercut values represent how effectively the model does not overcharge or over discharge the battery. Hard DQN model dominated the simpler models by more efficiently avoiding the battery over/short charge, as there were only 3 overcharge events and 4 undercut events, as opposed to the SOTA model which had 6 overcharge and 8 undercut events. This is an important statistic in the context of real-world usage when battery degradation becomes an essential concern.

The comparisons show that increasing the complexity of the DQN model is beneficial in terms of performance for all the evaluated ratios. The Hard DQN exceed performance of the simplest models as well as SOTA in terms of loss, accuracy, energy consumption and reward efficiency. In addition, the escalation curve demonstrates better battery health preservation and smoother decision making, which makes it as a viable solution for dynamic energy storage optimization. The state-of-the-art model, although very competitive, was not as fine-tuned, also on the forefront of reward, exploration vs exploitation trade-off, and smoothness in transitions. These insights emphasize the potential benefit of exploiting the iterative value estimation of deeper DQN architectures for better near-term and long-term performance of energy systems. This comparative study shows that DRL-based method, such as DQN, can effectively solve the DES system optimization problem, and more advanced DQN neural networks are worth exploring to obtain better performance in practice.

# conclusion

In this paper, we use the deep Q-networks (DQN) to control the operation of dynamic energy storage systems to balance energy storage and battery health to maximize the efficiency of the rewards. We judged the performance of the four experimental DQN architectures of easy level, intermediate-level, normal level and hard-level through different metrics related to the content of loss, accuracy, energy efficiency, exploration with exploitation and analysing the cumulated reward, reward efficiency, battery health preservation and transition smoothness for each DQN level. Results revealed that the deeper models (normal and hard) outperformed simpler ones in terms of both Accuracy and Cumulative reward value, which means better decision-making. Furthermore, balancing exploration-exploitation and reward efficiency was essential to reduce the overall energy consumption, whereas transition smoothness and battery health kept us informed about the model’s ability to infer transitions and its stability over longer horizons, respectively. The results emphasize that it is necessary to design adaptive and stable RL models to manage the dynamic environment and energy management tasks, and to guarantee the short-term performance and long-term sustainability for ES systems. These findings also lay a solid foundation for the future work of intelligent energy systems, which open an avenue for the further embedding of AI-based solutions in smart grids and renewable integration.

# Acknowledgments

The author would like to express sincere gratitude to Multimedia University and Universiti Teknikal Malaysia Melaka for providing the necessary facilities and support during the research process. Additionally, we extend our thanks to Universiti Teknologi Mara for their invaluable assistance in collaboration, idea generation, and dedicating time for discussions, which contributed significantly to the successful execution of this research project.

# References

1. M. Mnih, K. Kavukcuoglu, D. Silver, et al., *Human-level control through deep reinforcement learning* (Nature Publishing Group, London, 2015), pp. 486–491.
2. R. Kaur, D. Anand, U. Kaur, S. Verma, S.-W. Park, A. S. M. S. Hosen, and I.-H. Ra, “An advanced job scheduling algorithmic architecture to reduce energy consumption and CO2 emissions in multi-cloud,” Procedia Computer Science **57**, 545–553 (2015).
3. S. Baccari, M. Tipaldi, and V. Mariani, “Deep reinforcement learning for cell balancing in electric vehicles with dynamic reconfigurable batteries,” IEEE Transactions on Intelligent Vehicles (2024).
4. H. Xiao, X. Pu, W. Pei, L. Ma, and T. Ma, “A novel energy management method for networked multi-energy microgrids based on improved DQN,” IEEE Transactions on Smart Grid **14**(6), 4912–4926 (2023).
5. Y. Ding, X. Chen, and J. Wang, “Deep reinforcement learning-based method for joint optimization of mobile energy storage systems and power grid with high renewable energy sources,” Batteries **9**(4), 219 (2023).
6. S. Nyamathulla and C. Dhanamjayulu, “A review of battery energy storage systems and advanced battery management system for different applications: Challenges and recommendations,” Journal of Energy Storage **86**, 111179 (2024).
7. M. Murarka, P. R. Purohit, D. Rakshit, and A. Verma, “Progression of battery storage technology considering safe and sustainable stationary application,” Journal of Cleaner Production **377**, 134279 (2022).
8. S. Zhu, W. Lu, Y. Feng, and C. Sun, “Energy optimization for building energy management with thermal storage: A multi-agent deep reinforcement learning approach,” in Proceedings of the 43rd Chinese Control Conference (CCC) (IEEE, 2024), pp. 2805–2812.
9. H. Fan, E. Lu, W. Yu, L. Du, H. Wang, and D. Wang, “Multi-agent deep reinforced co-dispatch of energy and hydrogen storage in low-carbon building clusters,” IEEE Transactions on Network Science and Engineering (2023).
10. P. Balakumar, T. Vinopraba, and K. Chandrasekaran, “Deep learning-based real-time demand side management controller for smart building integrated with renewable energy and energy storage system,” Journal of Energy Storage **58**, 106412 (2023).
11. D. H. Nguyen, “The role of reinforcement learning control for optimizing building energy management systems” (2024).
12. D. Domínguez-Barbero, J. García-González, M. A. Sanz-Bobi, and E. F. Sánchez-Úbeda, “Optimising a microgrid system by deep reinforcement learning techniques,” Energies **13**(11), 2830 (2020).
13. M. W. Khan, J. Wang, M. Ma, L. Xiong, P. Li, and F. Wu, “Optimal energy management and control aspects of distributed microgrid using multi-agent systems,” Sustainable Cities and Society **44**, 855–870 (2019).
14. P. Mishra and G. Singh, “Energy management systems in sustainable smart cities based on the internet of energy: A technical review,” Energies **16**(19), 6903 (2023).
15. K. Shehzadi, T. Abbas, A. Zainab, and H. Li, “A survey of intruder detection in smart grid systems: Comparative analysis of rule-based, machine learning and deep learning,” International Journal of Security and Networks **20**(2), 67–92 (2025).