

Deep Q-Network-Based Energy Management of Battery-Supercapacitor Hybrid Energy Storage in a Renewable Microgrid

MinhPhong Le

Faculty of Electrical and Electronic, Thu Duc College of Technology, VietNam

Corresponding Author: MinhPhong Le, minhphongle232@gmail.com

ARTICLE INFO

Keywords: Deep Q-Network, Energy Management, Hybrid Energy Storage System, Battery-Supercapacitor, Renewable Microgrid

Received : 25, August

Revised : 22, September

Accepted: 26, October

©2026 Le: This is an open-access article distributed under the terms of the [Creative Commons Atribusi 4.0 Internasional](https://creativecommons.org/licenses/by/4.0/).



ABSTRACT

The increasing variability of renewable generation creates challenges for energy management in microgrids equipped with hybrid energy storage systems. This study develops a Deep Q-Network (DQN)-based strategy for coordinated battery-supercapacitor operation. The controller uses a seven-dimensional state representation and 25 discrete control actions within a reinforcement-learning environment. A rule-based controller (RBC) serves as the benchmark. During 50 training episodes, the cumulative reward improves, although the training loss remains variable. The evaluation results show distinct battery and supercapacitor utilisation patterns, but intermittent grid imports and negative cumulative reward indicate that the learned policy requires further improvement. The findings demonstrate the feasibility of DQN-based hybrid storage control, rather than overall performance superiority over RBC.

INTRODUCTION

The rapid deployment of solar and wind generation has increased the importance of microgrids as a means of coordinating distributed energy resources at the local level. By enabling generation and consumption to be managed within a defined electrical area, microgrids can contribute to improved energy efficiency, system resilience, and supply reliability (Mulaudzi et al. 2026). Their operating conditions, however, are strongly influenced by the variable nature of renewable generation. Solar irradiance and wind availability can change substantially over relatively short periods, while electricity demand and market prices may also vary with time. The resulting mismatch between renewable supply and consumption creates a need for energy management strategies that can respond effectively to changing operating conditions (Sargentis et al. 2026), (Dawn et al. 2025). Energy storage therefore represents an important component of renewable-rich microgrids because it can absorb surplus generation and provide energy during periods when local renewable production is insufficient (Dawn et al. 2025).

A hybrid energy storage system (HESS) offers an additional degree of flexibility by combining storage technologies with different electrical characteristics. In particular, batteries are capable of storing relatively large amounts of energy and are therefore appropriate for sustained energy balancing, whereas supercapacitors can deliver and absorb power rapidly and are better suited to short-duration fluctuations (Gowtham et al. 2025), (Ghoudelbourk, Yousfi, and Benbouhenni 2025). Coordinating these two storage technologies allows the HESS to exploit their complementary characteristics and can improve the flexibility of energy management in renewable microgrids (Srivastava et al. 2026a). Nevertheless, effective coordination cannot rely solely on the physical characteristics of the storage devices. The controller must continuously determine how the available storage resources should respond to renewable generation, load requirements, electricity prices, and their respective state-of-charge conditions (Eslami et al. 2026), (Banka and Kumar 2026; Ghoudelbourk, Yousfi, and Benbouhenni 2025).

Rule-based control (RBC) remains a commonly used approach for energy management because its operating logic is relatively simple and straightforward to implement. Typical rules specify charging when excess renewable energy is available and discharging when renewable production is insufficient to satisfy the load (Guo et al. 2026). Such deterministic strategies can provide satisfactory operation under predictable conditions; however, their predefined decision rules may become restrictive when the microgrid experiences rapidly changing generation, demand, or electricity prices. More advanced control and optimisation frameworks, including model predictive and mixed-integer approaches, have consequently been investigated to improve the adaptability of energy storage operation (Kavaliauskas et al. 2026), (Kırant-Mitić and Voss 2026). Deep reinforcement learning (DRL) provides another perspective by allowing a control agent to learn decision policies from repeated interactions with its operating environment rather than relying exclusively on manually specified rules (Ioannou et al. 2026), (Wang et al. 2026).

Motivated by these considerations, this study develops a Deep Q-Network (DQN)-based Energy Management System (EMS) for a renewable microgrid equipped with a battery-supercapacitor HESS. The proposed framework formulates energy management as a sequential decision-making problem in which the DQN agent observes the current operating state and selects an appropriate storage control action. The state representation incorporates renewable generation, load demand, electricity price, battery state of charge, supercapacitor state of charge, and temporal information. Through interaction with the simulated microgrid environment, the agent learns a control policy for coordinating battery and supercapacitor operation. The resulting strategy is evaluated against a conventional RBC approach under the same operating conditions.

The main contributions are threefold. First, a Gymnasium-compatible reinforcement-learning environment is formulated for coordinated battery-supercapacitor energy management. Second, a DQN controller with a seven-dimensional state representation and 25 discrete control actions is developed. Third, the learned policy is evaluated against rule-based control in terms of grid interaction, battery utilisation, supercapacitor utilisation, and cumulative reward. The study focuses on demonstrating the learning capability and operating behaviour of DQN-based HESS management rather than claiming overall performance superiority over the benchmark controller.

LITERATURE REVIEW

The increasing penetration of renewable energy has made effective energy management an important requirement for renewable-rich microgrids. Variations in renewable generation, load demand, and electricity prices require control strategies that can continuously adapt storage operation to changing system conditions. Recent reviews have highlighted the increasing use of advanced optimization, data-driven methods, and intelligent control for microgrid energy management (Atef et al. 2026; Tasmant et al. 2025). Energy storage is particularly important because it can absorb excess renewable generation and supply energy during periods of insufficient renewable production (Dawn et al. 2025; Srivastava et al. 2026), (Ton et al. 2026).

Hybrid energy storage systems (HESSs) provide additional flexibility by combining storage devices with complementary characteristics. Batteries are suitable for relatively long-duration energy management, whereas supercapacitors can respond rapidly to short-term power fluctuations. Their coordinated operation can therefore improve the flexibility of storage-based energy management (Gowtham and Suresh 2025; Ghoudelbourk, Yousfi, and Benboughenni 2025). Recent studies have investigated hierarchical, data-driven, and optimization-based approaches for HESS control, including PSO-based energy management, MPC, MILP, and other predictive strategies (Banka and Kumar 2026; Kavaliauskas et al. 2026a). Rule-based control remains attractive because of its simple implementation, but predefined rules may become restrictive under highly variable operating conditions (Guo et al. 2026).

Reinforcement learning (RL) and deep reinforcement learning (DRL) have consequently attracted increasing attention for adaptive energy management because control policies can be learned through interaction with the operating environment rather than being completely predefined. Recent studies have applied DRL to microgrid energy management, state-of-charge coordination, and renewable energy storage control (Ioannou et al. 2026; Khan et al. 2026; Li, Long, and Tian 2026) (However, the application of DQN specifically to the coordinated operation of battery-supercapacitor HESSs in renewable microgrids remains relatively focused on discrete control decisions and requires further investigation of how the learned policy redistributes the operating roles of the two storage technologies. This motivates the present study, which develops a DQN-based energy management framework with a seven-dimensional state representation and 25 discrete battery-supercapacitor control actions. The learned policy is evaluated against a rule-based controller under the same operating conditions, with emphasis on grid interaction, battery utilisation, supercapacitor utilisation, and cumulative reward.

Comparative Review of Recent Energy-Management Approaches

Recent research has investigated optimisation-based, predictive, and reinforcement-learning approaches for energy management in microgrids and hybrid energy storage systems. These approaches differ in their dependence on system models, decision-space formulation, computational requirements, and evaluation procedures. Table 2 summarises representative studies relevant to the present work. The comparison focuses on each study's principal method and research scope; details that are not explicitly reported in the respective publications should not be inferred.

The literature indicates that DQN provides a natural framework for problems formulated with discrete actions, while optimisation-based methods such as model predictive control (MPC), mixed-integer linear programming (MILP), and particle swarm optimisation (PSO) offer alternative ways to determine operating decisions. The present study focuses on formulating a DQN controller for coordinated battery-supercapacitor operation using a seven-dimensional state representation and 25 discrete action combinations. Its reported evaluation examines the resulting grid interaction and storage-utilisation patterns relative to an RBC benchmark. The current contribution is therefore the formulation and behavioural assessment of this particular DQN-based HESS controller, rather than a claim that DQN is universally superior to optimisation-based or conventional methods.

Table 1. Comparison of representative energy-management studies

Study	Method	System focus	Dataset/ environment	Evaluation or benchmark	Relevance to this paper
Ghoudelbourk, Yousfi, and Benboughenni (2025)	Energy-management strategy	Autonomous hybrid renewable power system	Renewable power-system study	See original study	Renewable-system energy management

Banka and Kumar (2026)	Hierarchical data-driven and PSO-based approach	HESS in DC microgrids	DC-microgrid environment	See original study	Optimisation-based HESS management
Kavaliauskas et al. (2026a)	MPC and MILP	Hybrid energy system	Hybrid-system model	MPC/MILP framework	Predictive and mathematical optimisation
Kavaliauskas et al. (2026b)	Deep reinforcement learning	Hybrid renewable energy system with storage	Study-specific simulation environment	See original study	DRL for storage control
Ioannou et al. (2026)	Deep reinforcement learning	Mini-grid SoC balancing	Mini-grid study	See original study	Learning-based storage coordination
Khan et al. (2026)	Deep reinforcement learning	Off-grid renewable microgrid with dual-battery storage	Study-specific simulation environment	See original study	DRL in renewable microgrids
Asadi, Eskandari, and Mansouri (2026)	Integrated optimisation and DRL	Battery energy storage	Study-specific BESS environment	See original study	Dynamic-aware battery energy management
Present study	DQN	Battery-supercapacitor HESS	Hybrid Energy Storage Dataset and a Gymnasium-compatible environment	RBC	Seven-dimensional state and 25 discrete battery-supercapacitor actions

METHODOLOGY

Dataset and Preprocessing

The study uses the Hybrid Energy Storage Dataset obtained from the Kaggle repository. The dataset includes time-series information related to solar and wind generation, load demand, electricity price, and battery-supercapacitor operating conditions. Solar and wind generation are aggregated into a renewable-power variable, while cyclical time features are constructed from the timestamp information.

The DQN configuration uses 50 training episodes. The reported controller evaluation covers 1,000 time steps. The total number of raw data samples, the temporal resolution, and the separation between training and testing data must be established from the original dataset and implementation before the evaluation can be described as an independent unseen-data test.

Aggregated Renewable Power

Solar and wind generation are represented by a single renewable-power term to simplify the representation of renewable availability at each decision interval. The aggregated renewable power is defined as

$$P_{RE}(t) = P_{solar}(t) + P_{wind}(t) \quad (1)$$

where, $P_{RE}(t)$ is the total renewable power available at time t , $P_{solar}(t)$ solar power generation at time t , $P_{wind}(t)$ is the wind power generation at time t , and t is the step within the simulation period.

This aggregation provides the DQN agent with a unified representation of renewable availability while retaining the combined contribution of the two renewable sources.

Cyclical Encoding of Time

Renewable generation and electricity demand exhibit recurring daily patterns. A direct numerical representation of the hour can introduce an artificial discontinuity between the last and first hours of a day. To avoid this issue, the hour of the day is transformed into two periodic components:

$$\sin_{hour} = \sin\left(\frac{2\pi h}{24}\right) \quad (2)$$

$$\cos_{hour} = \cos\left(\frac{2\pi h}{24}\right) \quad (3)$$

where, h is the hour of the day (0 - 23), π is a mathematical constant (3.14159), $\sin_{hour}$ is the sine representation of hourly periodicity, and $\cos(hour)$ is the cosine representation of hourly periodicity.

Min-Max Normalization

To obtain numerically consistent input variables, the continuous features are normalized using the Min-Max

$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

where, X is the original feature value, x_{min} is the minimum value of the feature, x_{max} is the maximum value, and X_{norm} is the normal feature value.

The transformation maps the selected continuous variables to the range [0, 1]. This scaling is particularly useful for neural-network-based reinforcement learning because it prevents differences in the numerical magnitude of the input variables from disproportionately affecting the learning process.

State-Space Formulation

The interaction between the DQN agent and the microgrid is described through a seven-dimensional state vector. At time t , the state is defined as

$$S_t = [P_{prenew,t}, L_t, EP_t, SOC_{bat,t}, \sin(h_t), \cos(h_t)] \quad (5)$$

where, S_t is the state vector at time t , $P_{prenew,t}$ is the renewable power generation, L_t is the load demand, EP_t is the electricity price, $SOC_{bat,t}$ is the battery state of charge, $\sin(h_t)$ is the sine-based temporal feature, and $\cos(h_t)$ is the cosine-based temporal feature.

Reward Function

The reward function determines how the DQN evaluates its decisions during interaction with the microgrid. The formulation is designed to discourage operating conditions that are undesirable from the energy-management perspective, particularly excessive grid imports, storage-related penalties, and violations of the allowable SoC range.

The reward at time t is formulated as

$$R_t = -(\alpha G_t + \beta D_t + \gamma V_t) \quad (6)$$

where, R_t is the reward received at time t , G_t is the grid import penalty, D_t is the storage degradation penalty, V_t is the state-of-charge violation penalty, α is the weight assigned to grid import penalty, β is the weight assigned to degradation penalty, and γ is the weight assigned to SoC violation penalty.

The negative formulation means that undesirable operating conditions reduce the instantaneous reward. During training, the DQN therefore seeks actions that reduce the accumulated penalties over successive time steps. In this way, the reward formulation provides the learning agent with a quantitative objective for balancing grid interaction and storage operation while maintaining the storage operating constraints (Wang et al. 2026; Asadi, Eskandari, and Mansouri 2026).

DQN-Based Energy Management Framework

The energy-management problem is solved using a Deep Q-Network (DQN). The DQN treats the microgrid as a sequential decision-making environment. At each time step, the agent receives the state S_t , evaluates the available actions, and selects the action associated with the highest estimated action value (Dawn et al. 2025; Ghodelbourk, Yousfi, and Benboughenni 2025).

The learning process can be expressed conceptually as

$$S_t \rightarrow DQN \rightarrow A_t \rightarrow R_t, S_{t+1} \rightarrow DQN \text{ update}. \quad (7)$$

Here, S_t is the current state, A_t is the selected control action, R_t is the resulting reward, and S_{t+1} is the state observed at the next decision interval.

The agent progressively improves its control policy through repeated interactions with the microgrid environment.

Action-Space Definition

The HESS uses a discrete action space in which the battery and supercapacitor are controlled simultaneously. Five discrete power levels are assigned to each storage device, resulting in 25 possible combinations:

$$NA = 5 \times 5 = 25 \quad (8)$$

These combinations represent different charging, discharging, and idle operating conditions (Banka and Kumar 2026; Kavaliauskas et al. 2026b).

Q-Value Learning

For each state-action pair, the DQN estimates the expected cumulative reward. The target value follows the conventional Q-learning formulation:

$$y_t = R_t + \gamma \max_{a'} Q_{target}(S_{t+1}, a') \quad (9)$$

where R_t is the immediate reward, γ is the discount factor, and Q_{target} represents the target-network estimate of the future action value.

The neural network is subsequently updated to reduce the difference between the predicted Q-value and the target value. A Smooth L1 (Huber) loss is used for this purpose, providing a more robust learning objective than a purely quadratic loss when relatively large temporal-difference errors occur (oannou et al. 2026; Khan et al. 2026; Asadi, Eskandari, and Mansouri 2026).

Exploration Strategy

An ϵ -greedy strategy is used to balance exploration and exploitation during training. At the beginning of training, the exploration probability is set to 1.0, allowing the agent to investigate the available action space extensively. As training progresses, ϵ is reduced to 0.05.

The resulting strategy gradually shifts the learning process from exploratory actions toward actions selected according to the learned Q-values. This allows the DQN to discover potentially effective battery-supercapacitor operating combinations before increasingly exploiting the learned policy (Wang et al. 2026; Li, Long, and Tian 2026).

DQN Network Architecture and Training Configuration

The DQN receives the seven-dimensional state vector as its input. The network consists of two hidden layers, each containing 128 neurons, with ReLU activation. The output layer contains 25 units corresponding to the 25 possible battery-supercapacitor action combinations. The Adam optimizer is used with a learning rate of 0.001, and the discount factor is set to 0.99. Gradient clipping with a maximum norm of 5.0 is employed to control excessively large parameter updates. Experience replay is implemented using a memory capacity of 50,000 transitions. A mini-batch of 128 samples is used for network updates. Training begins after the specified initial threshold, while 2,000 warm-up steps are performed using random actions. The target network is updated every 1,000 steps (Khan et al. 2026; Kavaliauskas et al. 2026b).

Hyperparameter Selection

The DQN hyperparameters were selected as a practical baseline configuration for the seven-dimensional state space and 25 discrete battery-supercapacitor actions. The network consists of two hidden layers with 128 neurons per layer, providing a moderate function-approximation capacity for the defined control problem. The learning rate of 0.001 and discount factor of 0.99 were adopted as initial settings for temporal-difference learning, while the Adam optimiser and Smooth L1 (Huber) loss were used for network training.

Experience replay with a capacity of 50,000 transitions and a mini-batch size of 128 was configured to provide a pool of past interactions for network updates. The ϵ -greedy exploration schedule decreases from 1.0 to 0.05 to allow broad exploration during early training and greater use of learned action values as training proceeds. The target network is updated every 1,000 steps, and gradient clipping is applied with a maximum norm of 5.0. The 50-episode training duration is treated as an initial experimental setting rather than evidence that the policy has converged.

A systematic hyperparameter search was not conducted in the present study. Accordingly, the selected values should not be interpreted as globally optimal. The observed fluctuations in training loss and the negative cumulative reward during evaluation support the need for further investigation of training duration, learning parameters, exploration scheduling, and reward-function weights.

The complete configuration is summarized in Table 2 and is retained from the original experimental framework.

Table 2. DQN Architecture and Training Hyperparameters

Parameter	Value	Rationale
Input dimension	7	Matches the seven-dimensional state representation
Hidden layers	2	Baseline network architecture
Neurons per hidden layer	128	Provides moderate function-approximation capacity
Output actions	25	Five battery levels $\times$ five supercapacitor levels
Optimizer	Adam	Optimises the network parameters
Learning rate	0.001	Initial learning configuration
Discount factor (γ)	0.99	Weights future rewards in the Q-learning target
Loss function	Smooth L1 (Huber)	Reduces sensitivity to large temporal-difference errors
Replay memory	50,000 transitions	Stores transitions for replay-based updates
Mini-batch size	128	Determines the number of transitions per training batch
Exploration rate (ϵ)	1.0 $\rightarrow$ 0.05	Gradually shifts from exploration toward exploitation
Training episodes	50	Initial training duration; does not establish convergence
Target network update interval	1,000 steps	Specifies the target-network update schedule

Overall Solution Procedure

The complete procedure consists of data preparation, DQN training, and final controller evaluation. After preprocessing, the time-series data are partitioned chronologically into training and testing subsets. During training, the DQN interacts only with the training environment, stores the resulting transitions in replay memory, and updates its online network according to the selected learning configuration. The target network is updated at the specified interval.

After training, the learned network parameters are fixed. The final evaluation is then conducted on the testing interval without further network updates, replay-buffer updates, or hyperparameter adjustment. The RBC is

evaluated over the same testing interval. This procedure is intended to assess the learned policy on operating data that were not used for training and to provide a consistent comparison between the two controllers.

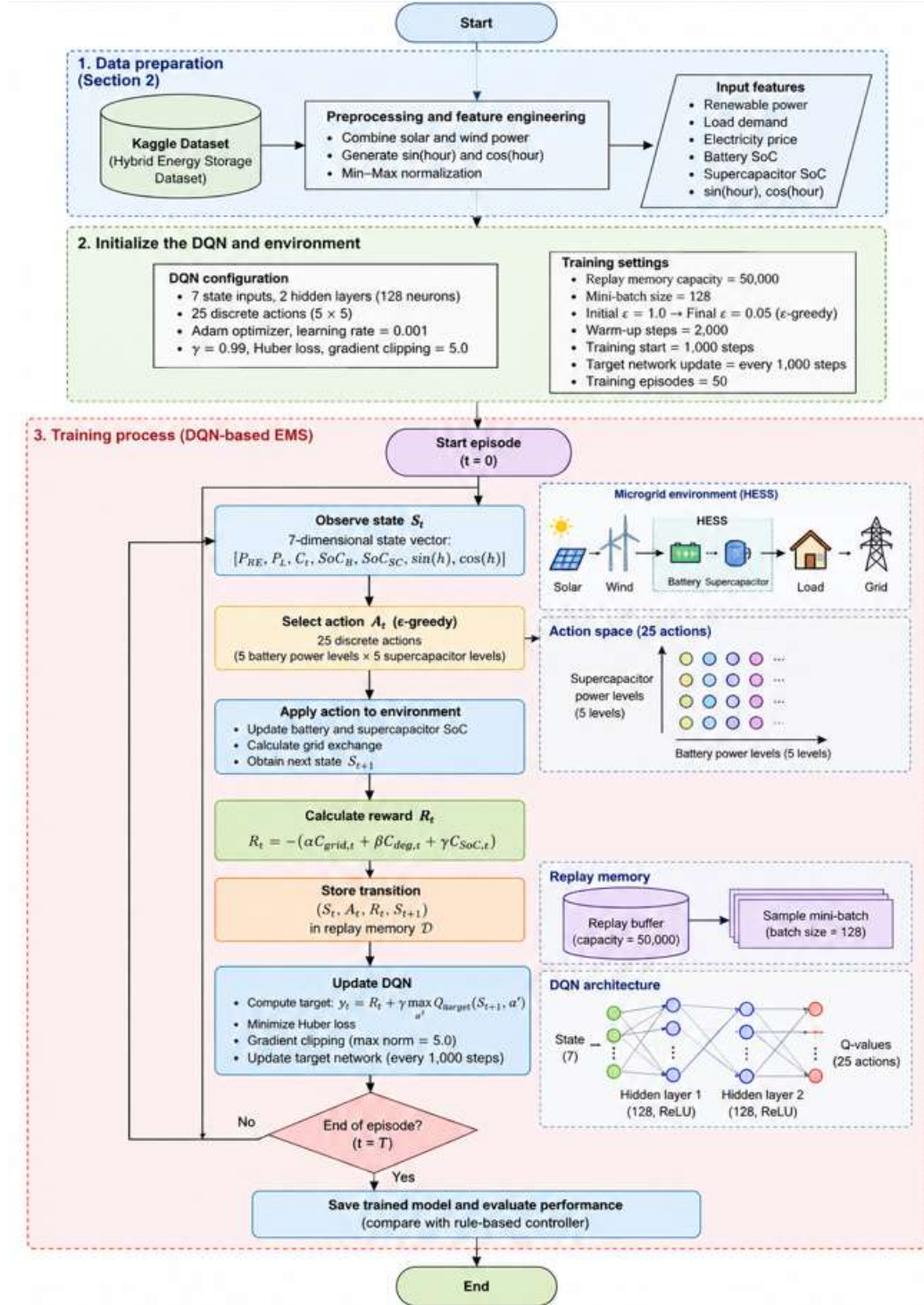


Figure 1. Overall workflow of the proposed DQN-based HESS energy management strategy

Figure 1 illustrates the overall workflow of the proposed DQN-based energy management methodology for the battery supercapacitor hybrid energy

storage system. The procedure begins with data preprocessing and feature construction, followed by initialization of the DQN agent and the microgrid environment. During training, the agent observes the seven-dimensional system state, selects one of 25 discrete battery-supercapacitor control actions using an ϵ -greedy strategy, receives the corresponding reward, and stores the transition in the replay memory. Mini-batch samples are then used to update the DQN parameters, while the target network is periodically updated to improve training stability.

RESEARCH RESULTS

Evaluation Protocol

The results are organised into two groups. The first group describes the learning process using the training reward and loss over 50 episodes. The second group evaluates controller behaviour using the designated testing interval. For a valid generalisation assessment, the trained DQN parameters must remain fixed during testing, and the testing observations must not be used for network updates or hyperparameter selection. The RBC and DQN must be evaluated using the same testing interval and the same underlying renewable-generation and load-demand conditions.

The training curves are used to describe learning behaviour and are not treated as evidence of generalisation. Grid interaction, battery utilisation, supercapacitor utilisation, and cumulative reward are reported separately for the testing period. This distinction prevents improvements observed during training from being interpreted as proof of superior performance on unseen operating conditions.

Renewable Generation and Load Characteristics

Figure 2 shows the renewable-generation and load profiles used to evaluate the energy management strategy. Solar generation follows a clear daily pattern, with negligible output during nighttime and higher production around midday. Wind generation exhibits greater short-term variability and provides a complementary renewable source. Load demand remains comparatively persistent and frequently exceeds the contribution of an individual renewable source. Such variability is characteristic of renewable-rich microgrids and represents one of the main motivations for adaptive energy-management strategies (Sargentis et al. 2026; Atef et al. 2026).

These operating characteristics create periods of renewable surplus and deficit, requiring coordinated charging and discharging of the battery-supercapacitor HESS. The resulting variability provides an appropriate operating environment for evaluating the ability of the DQN to adapt storage utilisation to changing generation and demand conditions.

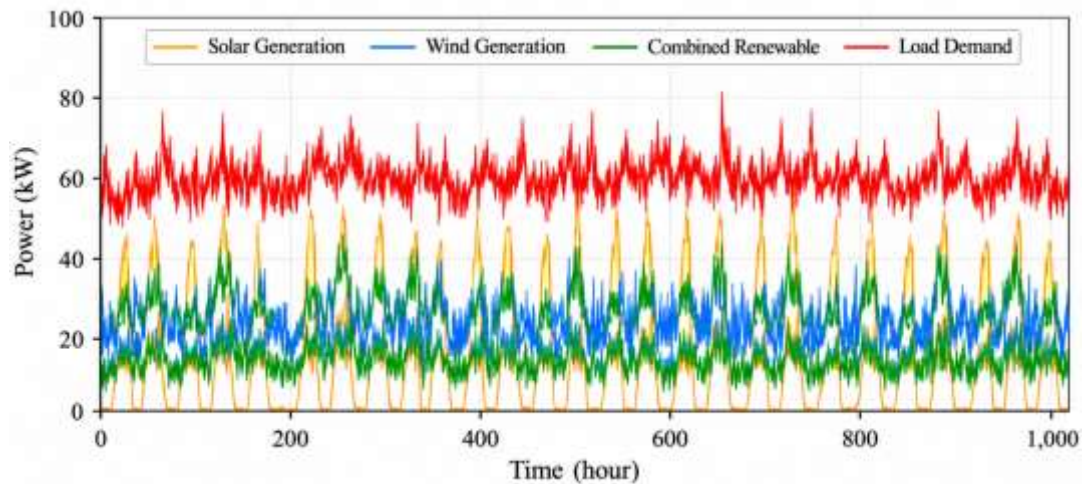


Figure 2. Renewable Supply and Load Demand

The complementary characteristics of the renewable sources are particularly relevant to HESS operation. Solar generation provides a relatively predictable daytime contribution, whereas wind generation introduces short-term fluctuations. Consequently, the energy management controller must simultaneously respond to predictable daily variations and stochastic changes in renewable availability.

DQN Training Performance

The DQN training behaviour is evaluated using cumulative reward and average training loss over 50 episodes. As shown in Figure 3, the cumulative reward improves from approximately -1200 during the initial training stage to approximately -400 in the final episodes, although fluctuations remain. This trend indicates progressive adaptation of the learned policy.

Figure 4 shows that the training loss initially reaches approximately 0.33 and subsequently decreases to below 0.05 within the first 10 episodes. The loss then fluctuates and reaches approximately 0.17 toward the final episode. The combined reward and loss trends indicate that the DQN learns meaningful control behaviour but does not achieve complete convergence within the 50-episode training period.

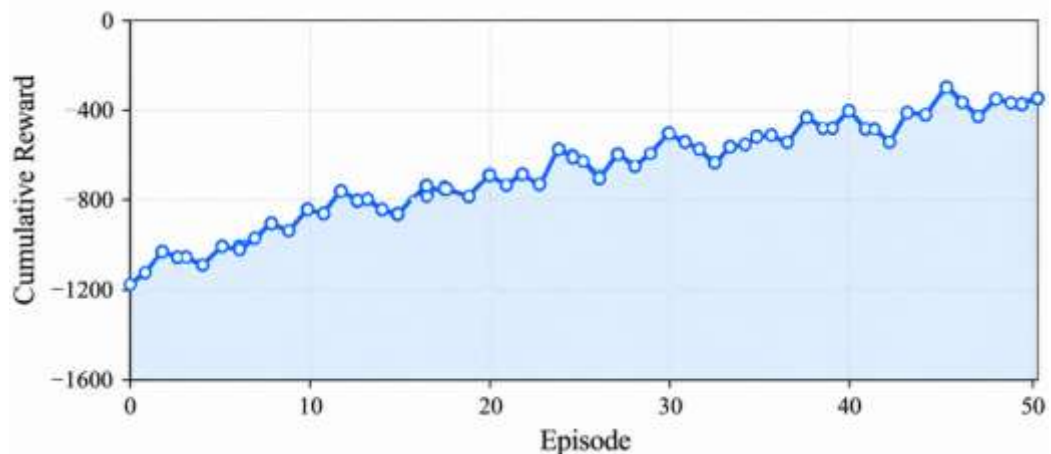


Figure 3. Training Reward per Episode

The improvement from approximately -1200 to -400 represents a substantial reduction in the accumulated penalty during training. More importantly, the trend indicates that the DQN agent is progressively modifying its control policy in response to the observed microgrid states rather than relying exclusively on random actions.

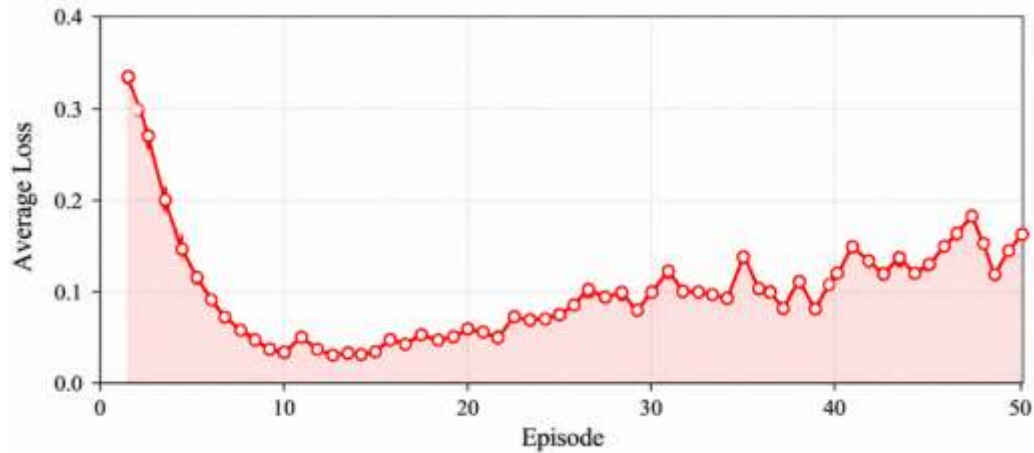


Figure 4. Training Loss per Episode

The training reward improves over the 50 episodes, while the training loss decreases during the early stage and subsequently fluctuates. These trends indicate that the agent updates its policy during training; however, they do not establish complete convergence or demonstrate generalisation to unseen operating conditions. The final policy must therefore be assessed separately on the testing data.

Rule-Based Control Behaviour

The rule-based controller (RBC) provides the benchmark against which the learned DQN policy is evaluated. Figure 5 shows the battery and supercapacitor behaviour under RBC control.

The battery begins at approximately 50 kWh and gradually decreases to approximately 10 kWh by the end of the simulation. This behaviour indicates that the RBC relies heavily on the battery for sustained energy balancing. The SC initially increases from approximately 0 kWh to around 5 kWh and subsequently remains relatively stable.

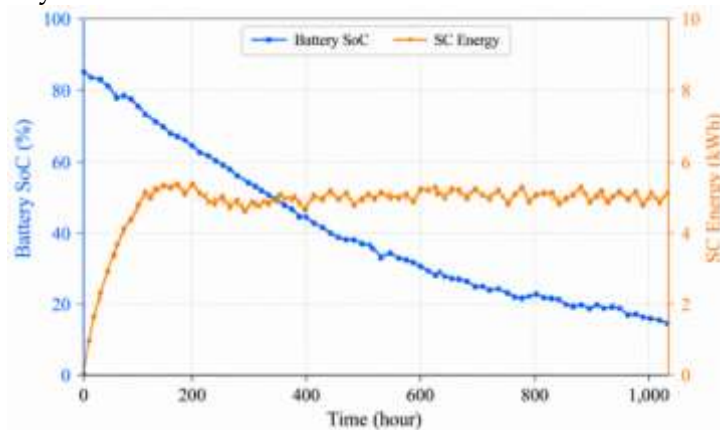


Figure 8 Battery and SC SOC under RBC

The RBC grid-import profile remains close to zero throughout the evaluated period. Thus, the benchmark controller primarily satisfies demand through local renewable generation and HESS operation. This provides a useful reference for assessing how the DQN redistributes the use of battery, SC, and grid resources.

Table 3. Observed RBC Operating Characteristics

Variable	RBC behaviour
Initial battery energy	50 kWh
Final battery energy	10–15 kWh, depending on the plotted evaluation interval
SC initial energy	0 kWh
SC subsequent level	5 kWh
Grid import	Close to 0 kW
Dominant storage role	Battery-oriented energy balancing

The RBC therefore exhibits a relatively deterministic operating pattern. Its main characteristic is maintaining a high degree of local energy autonomy, while the battery performs most of the long-duration balancing function.

Evaluation of DQN and RBC on Unseen Testing Data

The trained DQN and the RBC are evaluated over the same unseen testing interval. The DQN network parameters are kept fixed throughout evaluation. The comparison examines grid imports, battery utilisation, supercapacitor utilisation, and cumulative reward. These indicators are interpreted jointly because a change in one storage component's utilisation does not, by itself, establish an improvement in overall energy-management performance.

Grid Interaction

Figure 6 compares the grid-import profiles of the RBC and DQN strategies over 1,000 evaluation timesteps. The RBC remains close to zero throughout the evaluation, whereas the DQN produces intermittent grid-import events with observed peaks of approximately 0.85 kW.

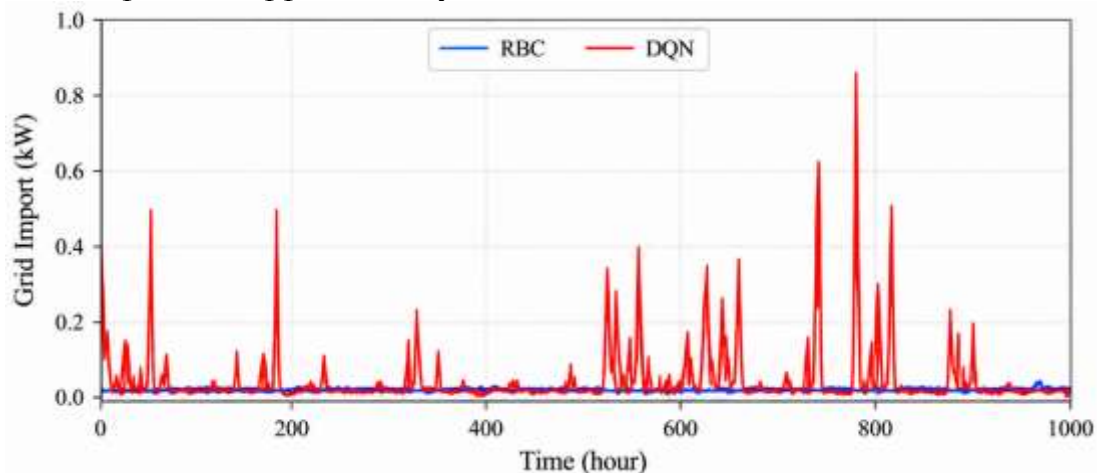


Figure 6. RBC and DQN Grid Imports

The difference indicates distinct resource-allocation behaviours. The RBC maintains near-zero grid interaction through its predefined control logic, while the DQN occasionally uses grid support when renewable generation and stored energy are insufficient to meet instantaneous demand. Under the current reward formulation, these grid-import events contribute negatively to the cumulative reward.

Table 4. Grid-Import Comparison

Controller	Observed grid-import behaviour	Maximum observed value
RBC	Consistently close to zero	0 kW
DQN	Intermittent fluctuations and spikes	0.85 kW

The grid-import profiles reveal a clear difference between the two control strategies. The RBC maintains grid imports close to zero under the reported operating conditions, whereas the DQN exhibits intermittent import events, with a maximum observed value of approximately 0.85 kW. These events contribute negatively to the reward under the current formulation. The result indicates that the learned policy has not yet reproduced the benchmark's near-zero-grid-import behaviour. Accordingly, the observed flexibility in resource allocation should not be interpreted as evidence of superior overall performance.

Battery Utilisation

Figure 7 compares the battery trajectories. Under RBC, the battery decreases from approximately 50 kWh to approximately 15 kWh over the evaluation period. In contrast, the DQN-controlled battery remains close to the lower operating region, fluctuating approximately between 0 and 2 kWh.

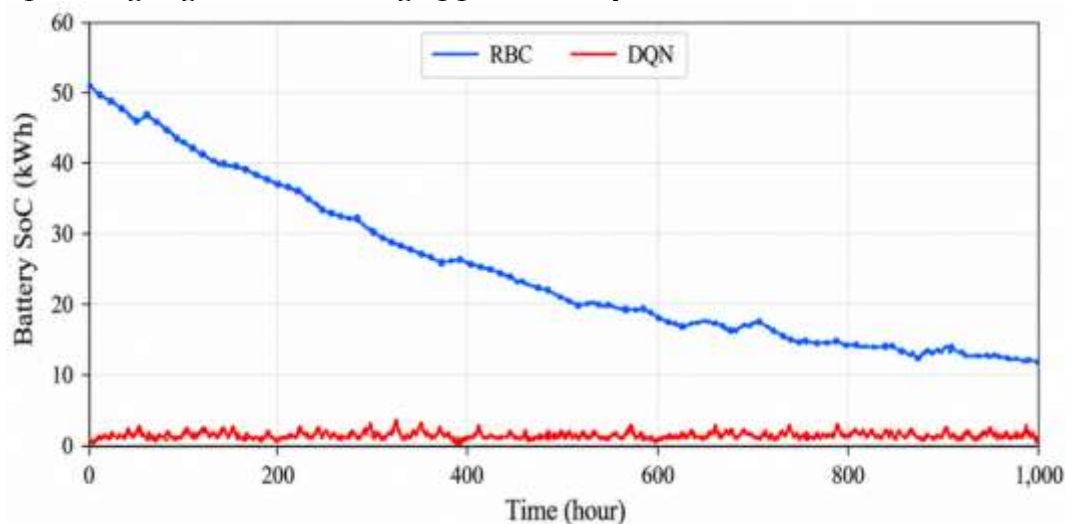


Figure 7. Battery SOC: RBC versus DQN

Table 5. Battery Operating Characteristics

Controller	Initial/typical level	Observed operating range/trend	Main behaviour
RBC	50 kWh	Declines to 15 kWh	Sustained battery utilisation
DQN	Low level	0–2 kWh	Limited battery cycling

The battery and supercapacitor trajectories show that the DQN adopts a different storage-utilisation pattern from the RBC. The DQN-controlled battery remains in a lower operating region, while the supercapacitor exhibits more frequent fluctuations. These observations indicate a redistribution of storage utilisation between the two devices. However, lower battery utilisation alone does not demonstrate reduced degradation or improved system efficiency, and more frequent supercapacitor operation does not necessarily imply better performance. Such conclusions require additional metrics, including energy throughput, SoC constraint violations, degradation estimates, and total grid interaction.

Supercapacitor Utilisation

Figure 8 shows the SC trajectories under the two control strategies. Under RBC, the SC remains relatively stable after its initial increase. The DQN, in contrast, produces substantially more frequent charging and discharging events.

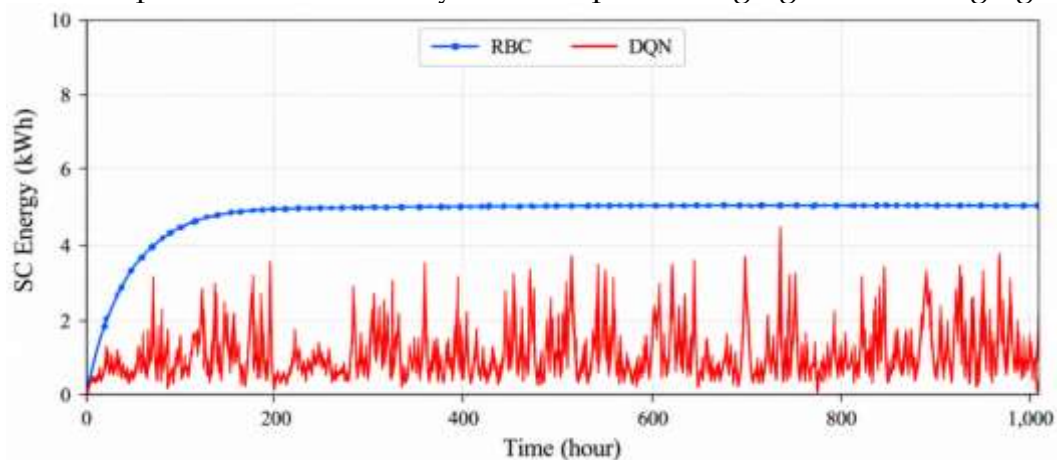


Figure 8. Supercapacitor SOC: RBC versus DQN

Table 6. Supercapacitor Operating Characteristics

Controller	Observed behaviour	Control characteristic
RBC	Rapid initial increase followed by relatively stable operation	Limited dynamic utilisation
DQN	Frequent charging/ discharging and overall fluctuations	Active short-term power balancing

The difference is particularly important because the SC is the faster storage component of the HESS. The DQN appears to exploit this characteristic more

actively than the RBC. The frequent SC fluctuations indicate that the learned controller responds to changing renewable generation and load conditions by dynamically redistributing short-term power support.

This result provides the clearest evidence that the DQN has learned a control strategy that differs structurally from the benchmark controller: the battery is used less intensively, while the SC is operated more dynamically.

Cumulative Reward

Figure 9 presents the cumulative evaluation reward over 1,000 timesteps. The RBC remains relatively close to zero, whereas the DQN cumulative reward decreases progressively to approximately -1200.

This result is important when interpreting the adaptive behaviour observed in the previous sections. Although the DQN demonstrates dynamic battery-supercapacitor coordination, its negative cumulative reward indicates that the current policy still accumulates substantial penalties during evaluation. Therefore, the present results demonstrate the feasibility of DQN-based HESS control but do not establish overall performance superiority over the RBC benchmark.

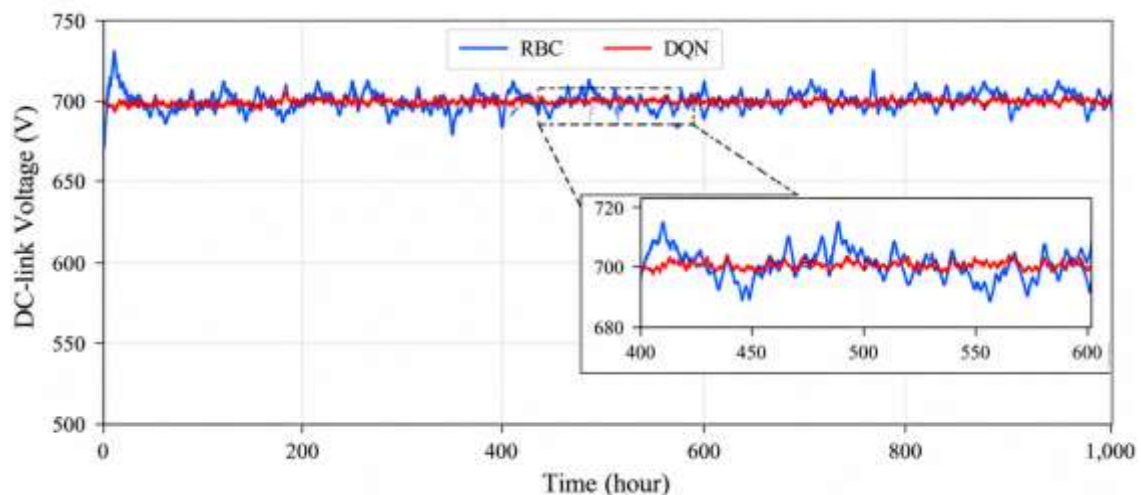


Figure 9. DC-Link Voltage Profiles under RBC and DQN Control

The DC-link voltage comparison provides an additional perspective on the operating behaviour of the two controllers. Nevertheless, voltage regulation should be assessed together with grid interaction, battery utilisation, supercapacitor utilisation, and cumulative reward. Considering these indicators jointly is important because a controller may produce a comparatively stable voltage profile while still incurring penalties in other aspects of the energy-management objective.

DISCUSSION

The results indicate that the DQN learns a storage-control policy during the 50-episode training process, as reflected in the improvement in cumulative training reward. Nevertheless, the fluctuations in training loss and the negative cumulative evaluation reward show that the current learning configuration has

not yet established consistently favourable operating performance Wang et al. 2026; Li, Long, and Tian 2026).

The comparison with RBC reveals a different allocation of storage resources. The DQN uses the battery less intensively and operates the supercapacitor more dynamically, but it also introduces intermittent grid imports. These outcomes illustrate why individual operating indicators should not be interpreted in isolation. Lower battery utilisation does not necessarily establish lower degradation, and more active supercapacitor operation does not by itself demonstrate better energy efficiency. A comprehensive assessment requires consistent evaluation of grid imports, storage energy throughput, state-of-charge constraints, reward or operating cost, and, where supported by the model, degradation-related indicators (Gowtham and Suresh 2025; Banka and Kumar 2026).

The present results should therefore be interpreted as a feasibility assessment of DQN-based battery-supercapacitor coordination rather than evidence of overall performance superiority over RBC. The current experimental comparison is limited to a rule-based benchmark, and a direct comparison with a stronger optimisation-based or alternative learning controller has not yet been established. Future experiments should implement MPC, MILP, PSO-based optimisation, or another RL controller using the same testing interval, system constraints, and evaluation metrics. Such a comparison would provide a more complete basis for assessing the relative benefits and limitations of the DQN approach (Asadi, Eskandari, and Mansouri 2026; Khan et al. 2026).

CONCLUSIONS AND RECOMMENDATIONS

This study developed a Deep Q-Network-based energy-management strategy for a battery-supercapacitor hybrid energy storage system in a renewable microgrid. The controller uses a seven-dimensional state representation and 25 discrete battery-supercapacitor actions to learn storage-control decisions through interaction with the simulated environment. The reported training results show an improvement in cumulative reward over 50 episodes, although the training loss remains variable. Compared with the rule-based controller, the DQN exhibits lower battery utilisation and more dynamic supercapacitor operation. However, intermittent grid imports and a negative cumulative evaluation reward indicate that the current policy has not yet demonstrated consistently favourable overall performance. The results therefore support the feasibility of learning-based battery-supercapacitor coordination, but do not establish the superiority of DQN over RBC. Further development should focus on improving the reward formulation, investigating the sensitivity of learning performance to hyperparameters, and evaluating the trained policy on independent testing data. A broader comparison with optimisation-based and alternative reinforcement-learning controllers is also required to establish the relative performance of the proposed approach.

ADVANCED RESEARCH

Future work should improve the DQN reward function and optimise its hyperparameters to enhance learning stability and control performance. The

trained controller should also be evaluated on unseen operating data under different renewable-generation and load conditions. Furthermore, comparisons with MPC, PSO-based optimisation, or alternative reinforcement-learning methods are needed to assess its performance more comprehensively.

ACKNOWLEDGMENT

The author sincerely thanks the Faculty of Electrical and Electronic Engineering, Thu Duc College of Technology, Viet Nam, for providing the academic environment and support for this study.

REFERENCES

- Asadi, Y., M. Eskandari, and M. Mansouri. 2026. "Revolutionising Battery Energy Storage Systems Energy Management: Dynamic-Aware Solutions with Integrated Quantum Particle Swarm Optimisation and Deep Reinforcement Learning." *IET Energy Systems Integration* 8 (1): 1–15. <https://doi.org/10.1049/esi2.70038>.
- Atef, M., S. Alahakoon, U. Mumtahina, P. Wolfs, T. Khatib, and M. Uddin. 2026. "A Review of Microgrid Energy Management Systems: Methods, Challenges, and Future Directions." *International Journal of Ambient Energy* 47 (1): 1–28. <https://doi.org/10.1080/01430750.2025.2595117>.
- Banka, S., and D. V. A. Kumar. 2026. "Hierarchical Data-Driven and PSO-Based Energy Management of Hybrid Energy Storage Systems in DC Microgrids." *Automation* 7 (2): 1–21. <https://doi.org/10.3390/automation7020050>.
- Dawn, S., S. Manideep, S. S. Rekha, C. R. Rao, K. D. Rao, A. Al Mansur, and T. S. Ustun. 2025. "Advancing Renewable Energy Integration in Deregulated Markets: The Role of Energy Storage, EVs, and Policy Frameworks." *Energy Exploration and Exploitation* 43 (6): 2745–2797. <https://doi.org/10.1177/01445987251348052>.
- Eslami, F., M. Gangineni, A. Ebrahimi, M. Rathnayake, M. Patel, and O. Lavrova. 2026. "A Review on Protection and Cybersecurity in Hybrid AC/DC Microgrids: Conventional Challenges and AI/ML Approaches." *Energies* 19 (3): 1–38. <https://doi.org/10.3390/en19030744>.
- Ghoudelbourk, S., R. Yousfi, and H. Benbouhenni. 2025. "Energy Management of an Autonomous Hybrid Renewable Power System." In *2025 IEEE 13th International Conference on Smart Energy Grid Engineering*, 190–194. IEEE. <https://doi.org/10.1109/SEGE65970.2025.11203682>.
- Gowtham, M., and S. Suresh. 2025. "Supercapacitor Integration in Hybrid Energy Storage Systems for Electric Vehicles: A Comprehensive Review." In *Proceedings of the 2025 International Conference on Intelligent Computing, Information and Control Systems*, 170–176. IEEE. <https://doi.org/10.1109/ICOIICS67115.2025.11390385>.
- Guo, Z., E. Grano, H. de Carvalho Pinheiro, and M. Carello. 2026. "A Comparative Evaluation of Rule-Based Strategies, ECMSs, and MPC Strategies for Fuel Cell Hybrid LCV Energy Management." *World Electric Vehicle Journal* 17 (3): 1–25. <https://doi.org/10.3390/wevj17030163>.
- Ioannou, I., S. Javaid, Y. Tan, and V. Vassiliou. 2026. "A Hierarchical Predictive-Adaptive Control Framework for State-of-Charge Balancing in Mini-Grids

- Using Deep Reinforcement Learning." *Electronics* 15 (1): 1–47. <https://doi.org/10.3390/electronics15010061>.
- Kavaliauskas, Ž., M. Milieška, G. Blažiūnas, G. Gecevičius, and H. Zhairabany. 2026a. "Optimisation of Hybrid Energy System Control Using MPC and MILP." *Applied Sciences* 16 (8): 1–22. <https://doi.org/10.3390/app16083690>.
- Kavaliauskas, Ž., M. Milieška, G. Blažiūnas, G. Gecevičius, and H. Zhairabany. 2026b. "Optimisation of Control for a Hybrid Renewable Energy System with Energy Storage Using Deep Reinforcement Learning Methods." *Sustainability* 18 (11): 1–43. <https://doi.org/10.3390/su18115443>.
- Khan, W., F. Renhai, A. Aziz, M. Z. Yousaf, Z. Cai, M. U. Iqbal, and M. S. Geremew. 2026. "Deep Reinforcement Learning-Based Energy Management for Design and Control of Off-Grid Renewable Microgrids with Dual-Battery Storage." *Energy Exploration and Exploitation* 44 (2): 821–869. <https://doi.org/10.1177/01445987251387368>.
- Kirant-Mitić, T., and K. Voss. 2026. "A Rule-Based Predictive Control Framework for Market and Grid-Oriented Operation in Thermally Activated Buildings." *Journal of Building Performance Simulation* 19 (2): 209–238. <https://doi.org/10.1080/19401493.2025.2565652>.
- Li, Z., W. Long, and H. Tian. 2026. "Current Applications and Future Prospects of Deep Reinforcement Learning in Energy Management for Hybrid Power Systems." *Energies* 19 (9): 1–50. <https://doi.org/10.3390/en19092216>.
- Mulaudzi, S. T., T. Rasimphi, and D. Tinarwo. 2026. "Assessing the Contribution of Microgrids to Sustainable Agriculture and Community Development at the Masia Development Centre." *Clean Energy Technologies* 2 (1): 1–16. <https://doi.org/10.54963/cet.v2i1.1945>.
- Sargentis, G., N. Papadodimas, I. Benekos, N. M. Katsoulakos, P. Dimitriadis, N. Tepetidis, and D. Koutsoyiannis. 2026. "Stochastic Assessment of Renewable Energy Reliability: A Case Study of North Euboea, Greece." *Journal of Energy and Power Technology* 8 (2): 1–13. <https://doi.org/10.21926/jept.2602008>.
- Srivastava, A., S. K. Jha, D. Gupta, and S. Singh. 2026. "A Critical Review of Microgrid Technologies: Storage and Control Systems with Effective Optimisation Techniques for Modeling Wind, Solar, Biomass, and Battery Energy Storage Systems." *Cureus Journals* 3 (1): 1–23. <https://doi.org/10.7759/s44388-025-00041-y>.
- Tasmant, H., B. Bossoufi, C. Alaoui, and P. Siano. 2025. "A Review of Machine Learning and IoT-Based Energy Management Systems for AC Microgrids." *Computers and Electrical Engineering* 127: 1–23. <https://doi.org/10.1016/j.compeleceng.2025.110563>.
- Ton, T. N., P. M. Le, and T. M. Le. 2026. "Multi-Objective Optimization of Distributed Generation Placement and Sizing in Active Distribution Networks Considering Harmonic Distortion." *International Journal of Electrical and Computer Engineering* 16 (2): 598–607.
- Wang, X., H. Zhang, H. Liu, and F. L. Lewis. 2026. "Control-Oriented Reinforcement Learning: A Survey of Recent Progress and Applications." *International Journal of Robust and Nonlinear Control* 36 (2): 694–718. <https://doi.org/10.1002/rnc.70152>