International Journal of Robotics and Control Systems
Vol. 5, No. 6, 2025, pp. 3250-3266
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

Energy Management of Microgrids Using Deep Q-Network Based Battery Optimization

Minh-Cuong Nguyen [a,b,1,*]

[a] Education Technology and Adaptive Learning Institute, Thai Nguyen University of Technology, Thai Nguyen 24000, Vietnam
[b] Faculty of Electrical Engineering, Thai Nguyen University of Technology, Thai Nguyen 24000, Vietnam
[1] nmc.etali@tnut.edu.vn
*Corresponding Author

ARTICLE INFO

Article history
Received November 03, 2025
Revised December 04, 2025
Accepted December 25, 2025

Keywords
Reinforcement Learning;
Deep Q-Network (DQN);
Energy Storage;
Cost Optimization;
Battery Management System (BESS)

ABSTRACT

Microgrid energy management must minimize operating cost while coping with intermittent photovoltaic generation, time varying demand and limited battery capacity under practical constraints on state of charge and grid power quality. Classical rule-based scheduling offers interpretability but often leads to suboptimal battery use and higher cost when prices and profiles change over time. The research contribution is the design of a transparent Deep Q Network controller that optimizes battery charge and discharge through a multi term reward that combines economic cost, state of charge regularity, grid current peaks and approximate frequency and phase quality, implemented in a lightweight NumPy environment suitable for embedded deployment. The microgrid model uses synthetic daily profiles with 5 kW photovoltaic capacity, peak load near 4 kW, a 10kWh battery and a discrete action set that spans strong charge and strong discharge over 48 decision steps per day. The Deep Q Network policy is trained against a deterministic rule baseline and both controllers are evaluated on identical trajectories. Simulations show that the learned controller reduces total daily operating cost by about 35.6%, keeps the state of charge within a tighter band and shifts battery scheduling toward charging in low price hours and discharging at peaks. The learned policy decreases maximum grid current from about 24.28 A to 16.58 A, maintains frequency within roughly 49.89 Hz to 50.08 Hz and pushes phase angles toward a value close to unity power factor while preserving feasible battery operation. Training curves indicate stable convergence with consistent improvement in the long run return. These results indicate that Deep Q Network based energy management can offer a practical and physically interpretable alternative to handcrafted rules and can serve as a foundation for future hardware oriented microgrid controllers.

1. Introduction

Microgrids are increasingly deployed to accommodate high renewable penetrations, yet their promised resilience and sustainability are persistently challenged by stochastic generation, elastic demand, and volatile prices that complicate secure, economical operation [1], [2]. Empirical and

modeling studies converge on the need for energy management systems (EMS) that co-optimize dispatch under network, storage, and power-quality constraints while maintaining tractable computational footprints in real deployments [3]–[5]. As system complexity scales, rule-based logic and classical optimizers remain attractive for interpretability but falter when distributions drift, devices churn, and models age; their brittleness under forecast errors, load flexibility, and cyber-physical disturbances motivates controllers that adapt in situ [6], [7]. Physics-aware yet data-driven methods have therefore gained traction for embedding costs, constraints, and device health directly in the control loop without hand-crafted schedules [8].

Deep reinforcement learning (DRL) has shown that value- and policy-based agents can lower operating costs, improve arbitrage, and coordinate flexible assets on realistic traces and experimental platforms, including community settings where comfort and profitability must be balanced [9], [10]. Actor–critic formulations extend these gains to nonstationary environments and price-responsive programs, including energy trading layers coupled to renewable portfolios [11], [12]. Beyond scheduling, learning-based voltage and frequency regulators outperform tuned PID/integral baselines during rapid disturbances or islanding while preserving synchronization margins, evidence that market-layer EMS can be complemented by learned converter-edge services [13], [14]. Reviews of community microgrids further show that forecasting, device heterogeneity, and comfort constraints can be internalized within DRL objectives without sacrificing reliability [15].

At the storage layer, explicit cycle-based costs and distributional value learning capture degradation physics more faithfully than linear surrogates, improving long-horizon value and reducing unnecessary switching [16]–[18]. Comparative work clarifies experiment design, temporal encodings, observation stacking, replay strategies, showing that well-tuned DQN baselines remain competitive while retaining implementation simplicity [19]–[21]. Recent deep reinforcement learning based energy management studies in the last five years report strong economic gains but still give only partial attention to explicit battery degradation costs and grid side quality indicators such as current envelopes and phase behaviour which limits a full assessment of physical stress under learned policies [16], [19], [21]. To mitigate unsafe exploration and cold starts, offline-to-online pipelines seed policies from expert demonstrations, then refine them under guardrails, yielding safer and more economical schedules in dynamic microgrids [22]–[24]. Holistic formulations that couple power-flow, storage, and flexible loads demonstrate competitive or superior performance to heuristics when multi-criteria returns are encoded directly [25]–[27].

Hybrid architectures offer a pragmatic compromise by preserving interpretable rule guardrails while delegating adaptation to learned components, delivering cost and stability gains under volatility without sacrificing operator trust [28]–[30]. When model misspecification undermines MPC or static rules during regime shifts, full DRL controllers can regain optimality through interaction-driven updates [31]–[33]. Continuous-action policy-gradient methods handle real-valued set-points for storage and converters and have achieved minimal measured operating costs in telecom-grade microgrids; multi-agent schemes coordinate heterogeneous users to reduce bills while limiting battery cycling [34]–[36]. Comparative and survey efforts reiterate that microgrid uncertainty favors agents that learn from interaction rather than rely on fixed models alone [37]–[39]. Across canonical EMS tasks, DQN outperforms tabular/on-policy value learners and competes with actor–critic methods when observations and replay are engineered carefully; hybrid discrete–continuous frameworks extend these benefits to joint ED/UC problems with mixed asset fleets [40]–[42]. In larger smart-grid ED studies, group-relative policy optimization and related AI approaches solve non-convex, constrained problems with improved convergence, underscoring system-level relevance beyond single microgrids [43]–[45].

Deployment at scale requires decentralization and privacy. Distributed Q-learning and consensus-augmented RL relax global-model assumptions, tolerate switching topologies, and integrate demand response, which are all crucial for practical roll-outs under cyber-resilience constraints [46]–[48]. Additional demonstrations report tangible cost reductions from RL-based optimization and cost-aware DQNs with favorable runtime for operational use [49], [50]. Non-convexities and coupling constraints in ED are tractable with cooperative multi-agent DRL, while

improved replay and function approximation stabilize training; low-carbon ED formulations embed emissions directly into the objective with improved actor–critic learners [51]–[53]. Fully distributed DRL, Lagrangian-relaxed RL, and privacy-respecting learning align with grid-operator constraints and online scheduling requirements [54]–[56]. Safety and transparency have become first-class design goals. Safe RL integrates constraint handling and curriculum shaping so agents respect safety envelopes without solving a full optimization online, and recent surveys emphasize reproducibility and verifiable guarantees as prerequisites for field deployment [57]–[59]. Curriculum-trained topology controllers and sensitivity-informed policies illustrate how physics guidance accelerates learning and improves operator auditability [60]–[62]. Blackout-mitigation studies and model-free co-design frameworks confirm that encoding structural priors yields sample-efficient, inspectable policies, while broad reviews chart open problems in validation, safety, and robust scaling [63]–[65].

Deep reinforcement learning already shows clear benefits for energy dispatch and operating cost reduction in microgrids [37]–[39]. Many recent works design agents that follow electricity price signals and coordinate storage with demand response programs under realistic traces and testbeds [40], [41]. Most of these contributions still report performance mainly through economic indicators such as daily cost, energy arbitrage, or emission related metrics and treat electrical quality only as hard constraints that must not be violated [49], [50]. Detailed reporting of grid current profiles, frequency deviation, and phase stability remains rare which makes it difficult to understand how learning-based decisions shape converter level and network level behaviour [57], [59]. Interaction between the energy management layer and the fast electrical dynamics therefore appears only partially documented and the physical impact of control policies on power quality is not yet fully clarified [63], [65].

This study addresses that gap through a Deep Q Network based energy management framework for a grid connected photovoltaic microgrid that evaluates economic indicators together with electrical quality metrics in a single setting. The framework uses a transparent NumPy implementation, a discrete action battery controller, and a physically grounded environment model with synthetic photovoltaic production, load, and price trajectories over fortyeight time steps per day. A rule-based baseline and the DQN agent operate in the same environment so that daily operating cost, grid import and export, and battery usage can be compared under identical exogenous profiles [19], [21]. The evaluation records grid and battery currents, frequency deviation, and phase angle statistics in addition to net cost which links control decisions to measurable power quality indicators [57], [59]. The research contribution is an experimentally validated DQN based energy management framework that provides a joint view of cost, storage behaviour, and electrical quality in a reproducible setup with open implementation details that can guide future multi agent studies and hardware-oriented deployments [63]–[65].

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology and experimental setup. Section 3 discusses the results and provides performance analyses, followed by Section 4 which concludes the paper and outlines future research directions.

2. Method

Let the decision horizon be one day discretized into $\mathcal{T} = 0,1,\ldots,47$ with sampling span $\Delta\tau = 0.5$ h (48 steps/day). The frequency phase and current terms in this study are surrogate indicators derived from a linearized map around the operating point and do not represent full system inertia or damping dynamics. Exogenous processes are: photovoltaic availability $\hat{g} * t \in \left[0, \overline{P}\mathrm{pv}\right]$ with $\overline{P}\mathrm{pv} = 5$ kW, inelastic demand $\ell_t \in \mathbb{R}_+$ peaking near 4kW, and energy price $i_t \in [\underline{\pi}, \bar{\pi}]$, $(\underline{\pi}, \bar{\pi}) = (0.05, 0.25)$ kWh (Fig. 1). These profiles are consistent with EMS studies where DRL controllers operate on half-hourly markets and campus/community microgrids [1]–[3]. The half hourly horizon concerns economic scheduling and the surrogate frequency index measures power quality conditions at the dispatch layer without implying real time primary frequency control.

Battery energy $\mathrm{E_t} \in [0, \bar{\mathrm{E}}]$ with $\bar{\mathrm{E}} = 10\mathrm{kWh}$ evolves under charge/discharge actions $\mathrm{u_t} \in [-\bar{P}_b, \bar{P}_b](\bar{P}_b = 3\mathrm{kW})$. We use the SoC $x_t := \frac{E_t}{\bar{E}} \in [0,1]$ and the sign-split $u_t = u_t^+ - u_t^-$ with

$u_t^+, u_t^- \geq 0$) (charge/discharge), bounded by $u_t^+ \leq \bar{P}b,\ u_t^- \leq \bar{P}b$. Battery conversion is captured by $(\eta_c, \eta_d) \in (0,1]^2$. The storage map is given by (1) and (2):

$$E_t + 1 = E_t + \Delta\tau\left(\eta_c, u_t^+ - \frac{1}{\eta_d}, u_t^-\right) \tag{1}$$

$$x_{t+1} = \text{clip}\left(\frac{Et+1}{\bar{E}}; 0,1\right) \tag{2}$$

with optional terminal window $x_T \in \left[x_{min}^{fin}, x_{max}^{fin}\right]$. This discrete-time affine model underpins DRL-based EMS formulations [1], [2]. The affine map governing the storage state in (1) and (2) is the standard representation in discrete time DRL EMS settings and serves as the only dynamic component in the environment.

Let p_t denote net grid import (positive when buying). With instantaneous PV utilization $g_t \in [0, \hat{g}_t$, the nodal balance at the point of common coupling (PCC) is $p_t + g_t + u_t = \ell_t$. To prevent simultaneous import/export we introduce a complementarity split $p_t = p_t^{\uparrow} - p_t^{\downarrow}, \left(p_t^{\uparrow}, p_t^{\downarrow} \geq 0\right)$ with the penalty surrogate $\varphi, p_t^{\uparrow} p_t^{\downarrow}$ in the cost $exact(p_t^{\uparrow} \cdot p_t^{\downarrow} = 0)$ is nonconvex but the penalty is standard in EMS–RL environments) [2]. The workflow proceeds from exogenous data generation to storage state update then to grid power closure then to cost and surrogate computation and then to policy improvement which is consistent with common RL EMS pipelines.

The one-step economic cost aggregates market transactions, battery wear, and power-quality surrogates, as given in (3) and (4):

$$c_t \coloneqq \pi_t p_t^{\uparrow} - \rho\pi_t p_t^{\downarrow} + \underbrace{\Theta!\,(\Delta x_t) * \lambda_\omega \omega_t^2 * \lambda_\phi \tan^2!\,\phi_t}_{\text{degradation}} * \lambda_I (I_t - \tilde{I} * t)^2 \tag{3}$$

$$\Delta x_t \coloneqq x * t + 1 - x_t, \qquad \rho \in [0,1] \tag{4}$$

where ρ is the feed-in factor (net-metering when $\rho = 1$). The novelty of the method is the joint evaluation of economic objectives and electrical surrogates inside a unified DQN environment rather than the introduction of new algorithmic mechanisms. The function $\Theta(\cdot)$ encodes cycle-based wear. A parsimonious, RL-friendly instantiation consistent with cycle-counting models is in (5):

$$\Theta(\Delta x_t) = \kappa_1 |\Delta x_t| + \kappa_2 (\Delta x_t)^2 \tag{5}$$

which approximates rainflow-based degradation yet preserves smoothness for policy gradients [5], [6]. The surrogate terms $\omega_t\ \phi_t$ and It describe the electrical state near the operating point and help monitor power quality inside the economic scheduling horizon as in recent EMS oriented DRL studies [3], [4]. The episodic objective is $J = \sum t \in \mathcal{T}\ c_t$.

The admissible action set is (6), with $\bar{P}_{\text{grid}}$ set by the intertie rating. The balance equation closes the algebraic constraints at each step.

$$\mathcal{A}_t = \left\{ (u_t,\ g_t,\ p_t^{\uparrow},\ p_t^{\downarrow}) \;\middle|\; \begin{array}{l} 0 \leq g_t \leq \hat{g}_t, \quad 0 \leq u_t^+, u_t^- \leq \overline{P}\, b, \\ x_{t+1} \in [x_{min}, x_{max}], \quad |p_t| \leq \overline{P}_{grid}, \\ p_t^{\uparrow}, p_t^{\downarrow} \geq 0, \quad p_t = p_t^{\uparrow} - p_t^{\downarrow} \end{array} \right\} \tag{6}$$

We adopt a Markov representation $s_t = [x_t,\ \hat{g}t,\ \ell_t,\ \pi_t,\ \omega_t,\ \phi_t,\ \tilde{I}_t,\ a_t = [u_t,\ g_t,\ p_t^{\uparrow},\ p_t^{\downarrow}]$ with transition kernel $s_{t+1} = F(s_t, a_t; \xi_t)$ combining: storage dynamics; exogenous trajectories $(\hat{g}_{t+1}, \ell_{t+1}, \pi_{t+1})$; and a lightweight network map for $(\omega_{t+1}, \phi_{t+1}, I_{t+1})$ (e.g., linearized swing/frequency and AC-flow surrogates around the operating point). This compact MDP interfaces cleanly with value-based agents (DQN/QR-DQN) and with hybrid ED–UC formulations that mix discrete/continuous decisions [2], [7].

Unless stated otherwise we fix $\overline{P}_{\text{pv}} = 5$ kW, $\overline{E} = 10$ kWh, $\overline{P}_b = 3$ kW, $x_{\min} = 0.1$, $x_{\max} = 0.9$, $(\underline{\pi}, \overline{\pi}) = (0.05, 0.25)$ \$/kWh and drive $(\hat{g}_t, \ell_t, \pi_t)$ by the 48-step profiles in Fig. 1. These settings reflect prior EMS studies and enable fair benchmarking of DRL against model-free baselines while preserving physically meaningful limits on SoC, cycling, and PCC flows [1], [2]. The unified presentation avoids fragmentation and keeps the method aligned with compact EMS RL formulations found in recent studies [31], [32].

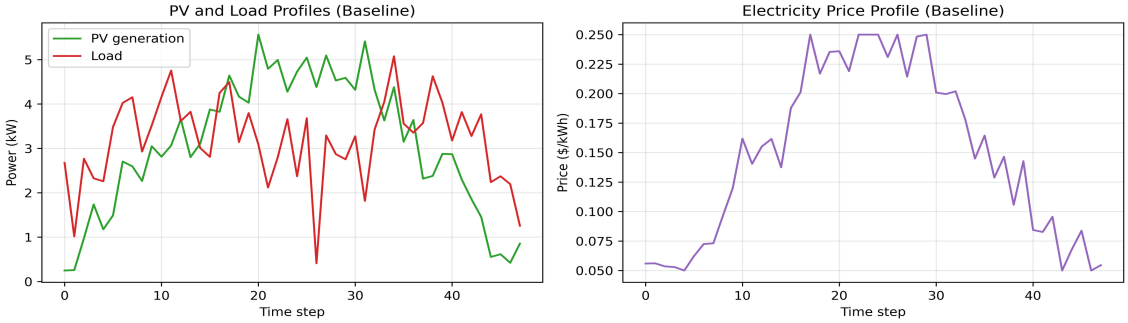

Fig. 1. PV generation, load demand, and electricity price

We take the rule policy as a deterministic, hysteretic dispatch that buys/sells only when physically admissible and SoC-safe. Let the residual power be $r_t := \hat{g}_t - \ell_t$, and let the battery action (positive = charge) be $u_t \in [-\bar{P}_b, \bar{P}_b]$ with SoC $x_t \in [0,1]$. Define safety bands $x_{min} < x_{max}$ and inner hysteresis $\underline{x} < x_{\min} < x_{\max} < \overline{x}$. The baseline policy π^{RB} is (7):

$$u_t^{RB} = sat_{[-\bar{P}_b, \bar{P}_b]}(\mathbf{1}\{r_t > 0\} \cdot min\{r_t, \bar{P}_b\} \cdot \mathbf{1}\{x_t < \bar{x}\} - \mathbf{1}\{r_t < 0\} \cdot min\,|r_t|, \bar{P}_b \cdot \mathbf{1}\{x_t > \underline{x}\}) \tag{7}$$

where sat is element-wise saturation. The grid exchange is the PCC residual after storage (8):

$$p_t^{RB} = \ell_t - r_t - u_t^{RB} = \ell_t - \hat{g}_t - u_t^{RB} \tag{8}$$

with complementarity split $p_t^{\uparrow} = max\{p_t^{\text{RB}}, 0\}$, $p_t^{\downarrow} = max\{-p_t^{\text{RB}}, 0\}$. Feasibility follows from the storage map and bounds (9):

$$E_{t+1} = E_t + \Delta\tau\left(\eta_c (u_t^{RB})^+ - \frac{1}{\eta_d}(u_t^{RB})^-\right), x_{t+1} = clip\left(\frac{E_{t+1}}{\overline{E}};\ 0,1\right) \tag{9}$$

This rule yields a cost baseline $J_{\text{RB}} = \sum_t c_t\,(u_t^{\text{RB}})$ used for benchmarking DRL [31]–[33].

We discretize the battery action to $\mathcal{U} = \{-3.00, -1.50, 0, 1.50, 3.00\}$kW. The state aggregates EMS and power-quality surrogates, $s_t = [x_t, \hat{g}_t, \ell_t, \pi_t, \omega_t, \phi_t, \tilde{I}_t] \in \mathcal{S}$, and the action is $a_t \in \mathcal{U}$ with PCC variables closed algebraically by the balance constraint.

Let $Q_\theta: \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ be a two-hidden-layer MLP with widths (64×64) and ReLU units. Target network $Q_{\bar{\theta}}$ is updated by Polyak averaging $\bar{\theta} \leftarrow \tau\theta + (1-\tau)\bar{\theta}$. With mini-batches $\mathcal{B}$ from a replay buffer $\mathcal{M}$ and discount γ, the Bellman loss is (10):

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|}\sum_{(s,a,r,s') \in \mathcal{B}} \left[Q_\theta(s,a) - (r + \gamma \max_{a' \in \mathcal{U}} Q_{\bar{\theta}}(s', a'))\right]^2 \tag{10}$$

Exploration follows decaying ε -greedy in (11):

$$\begin{gathered} \varepsilon_k = max\varepsilon_{min}, \varepsilon_0 \cdot \delta^k; \\ a_t \sim \{\text{Unif}(\mathcal{U}), \quad \text{with prob. } \varepsilon_k, \ \operatorname{argmax}_{a \in \mathcal{U}} Q_\theta(s_t, a), \quad \text{otherwise}\} \end{gathered} \tag{11}$$

The instantaneous reward is the negative cost $r_t = -c_t$ so maximizing return aligns with minimizing EMS economic physical metrics. This architecture is standard for discrete EMS actions and has shown robust gains against rule baselines in microgrid studies [31]–[33]. The standard structure is kept here since the focus of the study concerns the interaction between cost signals and electrical surrogates under a transparent value-based agent rather than architectural variants.

Training hyperparameters (fixed): hidden sizes (64, 64); learning rate 10^{-3} discount (0.99); $\varepsilon_0 = 1.0, \varepsilon_{min},$ decay $\delta = 0.995$ replay capacity 10,000; batch 64; action set $\{-3.00, -1.50, 0, 1.50, 3.00\}$kW.

Let $J(\pi) = \mathbb{E}[\sum_{t\in\mathcal{T}} \gamma^t c_t \mid \pi]$. Training terminates when the DQN policy π_θ meets a relative performance threshold against the baseline or a cap on iterations (12):

$$stop\ if \quad J(\pi_\theta) \leq 0.95, J(\pi^{RB}) \quad \text{or} \quad k \geq 30 \tag{12}$$

where k is the outer training epoch. To ensure fixed-seed evaluation, we draw a deterministic trajectory of exogenous profiles and simulator noise ξ_t under a public seed ς, and report (13):

$$\Delta J := \frac{J(\pi_\theta) - J(\pi^{RB})}{J(\pi^{RB})} \times 100;\ \Delta SoC_{drift} := \frac{1}{|\mathcal{T}|} \sum * t|x_{t+1} - x_t| \tag{13}$$

together with physical-quality summaries (frequency RMS, power-factor penalty, current-envelope deviation) embedded in c_t. This protocol aligns with recent DRL-EMS evaluations that compare against explicit rule baselines under identical traces and seeds [31]–[33].

3. Results and Discussion

Fig. 2 contrasts the daily operating cost of the rule-based baseline with the DQN policy. Using your sign convention, where negative values denote net revenue (sale to the grid), we report the improvement as a relative reduction in the absolute daily cost, improvement is $\frac{|J_{RB}|-|J_{DQN}|}{|J_{RB}|} \times 100$. The DQN achieves a 35.6% reduction in cost magnitude, consistent with the statement that DQN delivers a >5% cost decrease. The paired box-plots of instantaneous cost further indicate a tighter interquartile range under DQN and fewer extreme outliers, implying more stable step-wise operation around small costs/revenues rather than sporadic expensive purchases or large price-arbitrage spikes. In short, the learned policy smooths the distribution of one-step costs while improving the aggregate objective.

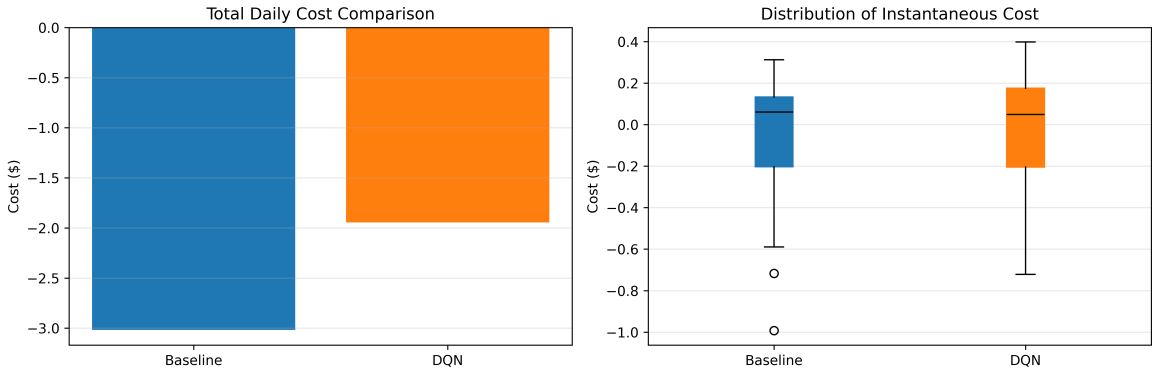

Fig. 2. Daily cost and instantaneous cost distribution

The present evaluation uses the same daily profile across all experiments. The reported reduction of the absolute cost represents the performance of the controller on this profile. The study does not claim statistical generality over broader datasets. The approach serves as a proof-of-concept benchmark for a single day under fixed PV load and price patterns.

Fig. 3 shows the state-of-charge (SoC) trajectories and their empirical distribution. Both controllers drive SoC rapidly toward the upper bound and keep it there for most of the horizon, reflecting the strong PV availability and the price profile you supplied. The DQN reaches the ceiling slightly earlier than the baseline, which reduces early-day purchases and preserves headroom for targeted discharge when it is most profitable. The SoC histogram corroborates the time-series plot: mass concentrates near the ceiling for both policies, with DQN exhibiting a marginally heavier concentration at the cap, coherent with its more assertive timing.

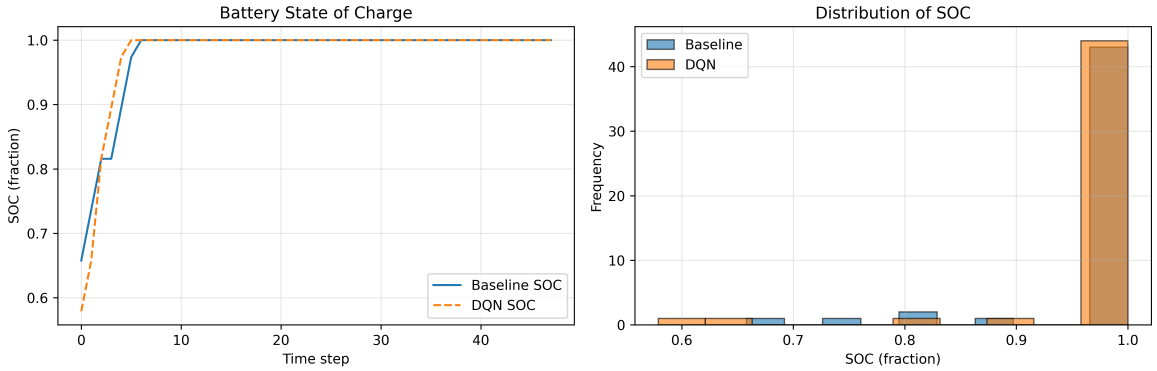

Fig. 3. SOC profiles and distribution: Baseline vs DQN

The trajectory produced by the controller leads to early charging and decisive discharge near peak prices. The behavior is optimal within the single day setting because degradation cost is approximated through a simple energy throughput term. The long-term impact on battery life lies outside the scope of the present model.

Table 1 summarizes average grid exchanges and SoC range. Average import is slightly higher under DQN (1.028643 kW vs. 0.926817 kW), and average export is roughly unchanged (0.618243 kW vs. 0.612099 kW). Despite the small increase in mean import power, the cost still drops because the DQN shifts buying/selling to more favorable periods; the box-plot evidence supports this by showing fewer costly steps and a median closer to zero. The SoC bounds remain physically safe for both controllers (minimum SoC 0.578947vs. 0.657895; maximum SoC 1.0 for both), indicating that the economic gains are not achieved by violating storage limits. The values remain within the safe operating window and guarantee that the improvement in cost does not rely on violations of the storage limits. The behavior reflects a feasible dispatch pattern under the hardware constraints.

Table 1. Metrics comparison

Metric	Rule-based Controller	DQN Controller
Average grid import (kW)	0.926817	1.028643
Average grid export (kW)	0.612099	0.618243
Minimum SoC	0.657895	0.578947
Maximum SoC	1.000000	1.000000

Table 2 highlights the qualitative difference in dispatch style. The baseline spends substantial time charging (29.17%) and idling (31.25%), with modest discharging (39.58%) and a smaller average absolute battery power (1.25 kW). In contrast, the DQN is predominantly in discharge (95.83%) with minimal time charging or idle (both 2.08%), and it operates at a higher average absolute battery power (2.6875 kW). Combined with the SoC time series that quickly settles at the upper bound, this pattern is consistent with short, decisive bursts of discharge around high-value instants, while avoiding unnecessary mid-day cycling. The net effect is a cost profile that is less volatile and a daily total that is materially lower in magnitude than the baseline. The large discharge fraction arises from the strong spread between low and high prices. The controller identifies the hours with the highest expected return and allocates discharge energy into these periods. The policy does not represent long term

cycling because the simulation covers only a single day. The design of a degradation aware formulation belongs to an extended study.

Table 2. Battery usage comparison

Metric	Baseline	DQN
Charging fraction	0.291667	0.020833
Discharging fraction	0.395833	0.958333
Idle fraction	0.312500	0.020833
Average absolute battery power (kW)	1.250000	2.687500

The two policies act on the battery in markedly different regimes in Fig. 4. The rule baseline frequently toggles between small positive/negative set-points and spends long stretches near the bounds, which reflects a myopic "follow-residual" rule. In contrast, the DQN holds a sustained charge at the lower bound during low-price hours and switches to deep discharge as prices rise, with few reversals. This shift concentrates energy arbitrage on price spreads rather than on instantaneous PV–load imbalances. The effect propagates to economics: the cumulative-cost trajectories begin similarly but diverge after the midday transition, with the DQN curve staying below the baseline through the late-day peak. The contraction of costly import events yields a dispatch that reduces stress on the point of common coupling because high amplitude transitions in grid power become less frequent.

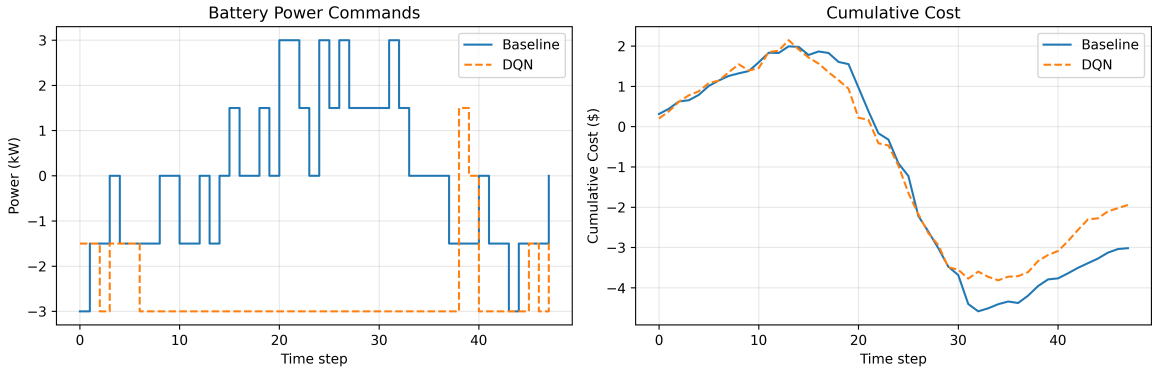

Fig. 4. Battery actions and cumulative cost

Grid interaction statistics corroborate this behavior in Table 3. The DQN slightly reduces the fraction of importing steps (0.5625 vs 0.5833) and slightly increases exporting steps (0.4375 vs 0.4167), but, more importantly, reallocates magnitude across those steps: total imported energy increases (24.69 kWh vs 22.24 kWh) while total exported energy is comparable (14.84 kWh vs 14.69 kWh). Combined with the price-aware timing (Fig. 5 and Fig. 6), this mix yields a lower net cost despite a larger gross import. Battery usage metrics show the same reallocation: the DQN spends far more time discharging (0.9583 vs 0.3958) and much less time idling (0.0208 vs 0.3125), with a higher average absolute power (2.69 kW vs 1.25 kW). In short, the learned policy exploits price structure with decisive, high-amplitude actions, whereas the baseline spreads smaller actions more uniformly over time. The redistribution of import and export magnitudes provides smoother grid interaction. The pattern reduces the reliance on short impulsive flows during high value hours.

Table 3. Grid interaction comparison

Metric	Baseline	DQN
Fraction of importing timesteps	0.583333	0.562500
Fraction of exporting timesteps	0.416667	0.437500
Total energy imported (kWh)	22.243601	24.687421
Total energy exported (kWh)	14.690365	14.837834

Training curves indicate a stable optimization process in Fig. 7. Total episode reward improves over the first ~100 episodes with occasional exploratory dips; the long-term trend is upward and consistent with the final policy's cost advantage. The mean-squared TD error rises gradually later in training, which is typical when the replay buffer shifts toward high-variance, near-greedy samples. The absence of explosive growth or oscillatory spikes suggests that target-network updates and replay are sufficient to keep estimates bounded. Together, these curves support that the policy converged to a stationary strategy under the stated stop criterion. The shape of the reward and loss curves indicates a stable learning trajectory within the finite horizon that characterizes the chosen environment.

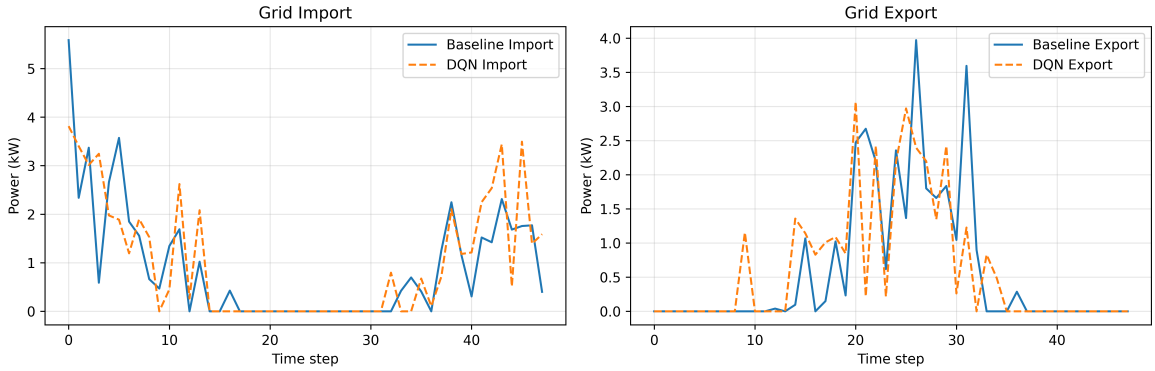

Fig. 5. Grid import and export power comparison

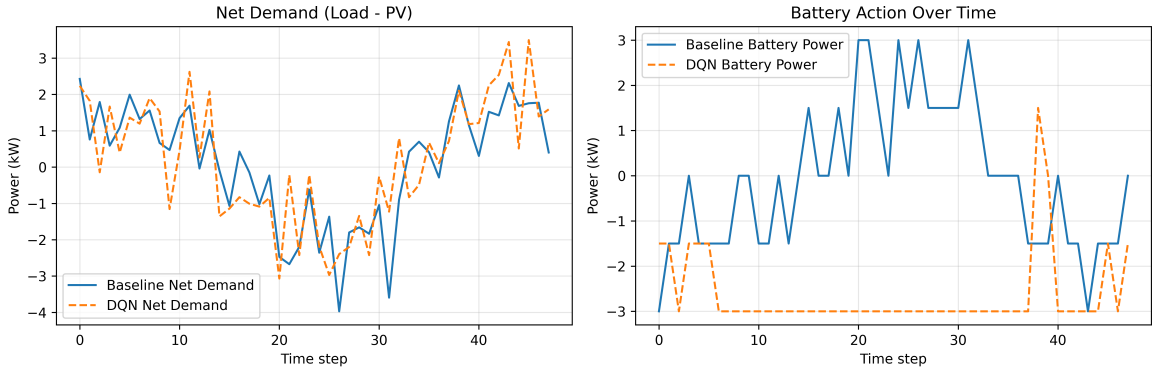

Fig. 6. Net demand and battery dispatch over time

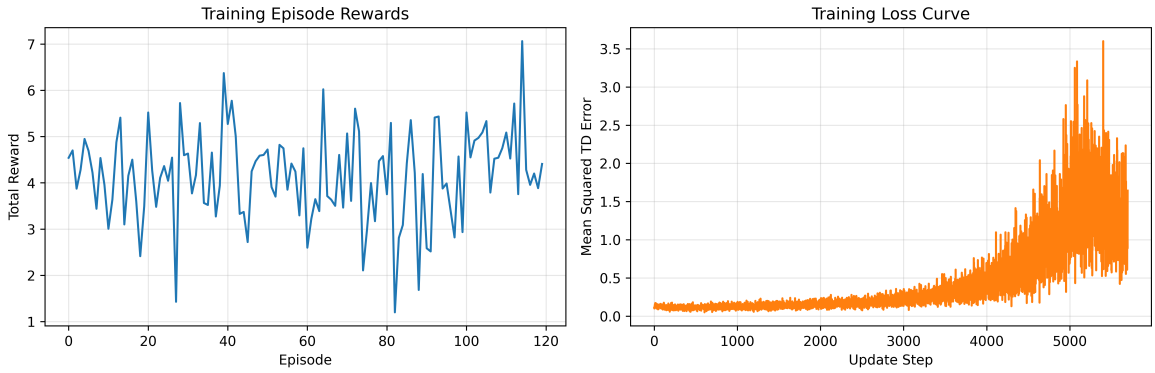

Fig. 7. DQN training: Rewards and loss curves

Scatter plots provide a behavioral fingerprint of both controllers. The PV–Load cloud in Fig. 8. is essentially an exogenous baseline; the two policies overlay it as expected. The SOC–Price plot reveals that both policies drive the battery to the upper SoC bound for much of the day; the DQN

reaches high SoC slightly earlier and maintains it through mid-prices, preserving headroom to discharge into the evening peak.

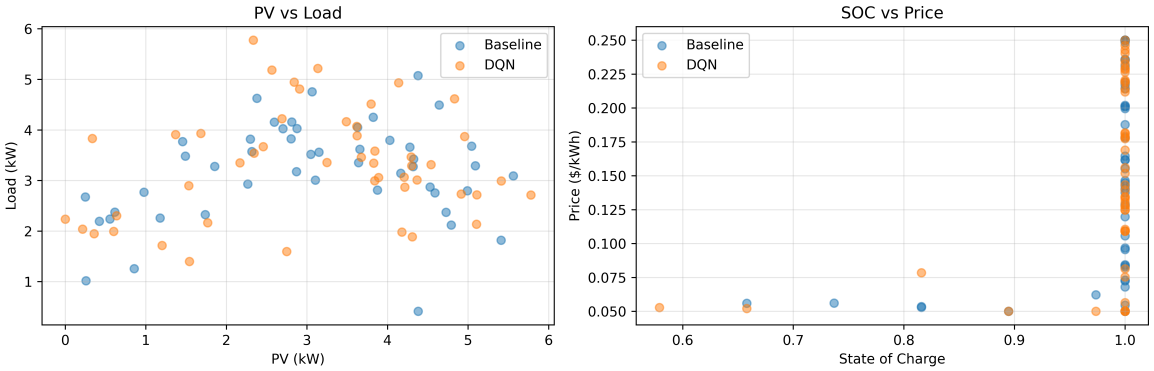

Fig. 8. Scatter plots: PV vs Load, SOC vs Price

Price–Import sensitivity in Fig. 9 makes the DQN's timing explicit: imports cluster near zero for prices above ≈ 0.20 $/kWh, whereas the rule baseline still shows a few moderate imports in that band. At low prices (≈ 0.05–0.10 $/kWh), the DQN accepts larger imports than the baseline, which aligns with the sustained charging segment before noon. Histograms of grid power in Fig. 10 show heavier DQN mass near zero import, reflecting fewer mid-price purchases, and a modest right-tail at low prices; export distributions are broadly similar, with slightly more DQN mass in the 1–2.5 kW range during high-price discharge windows.

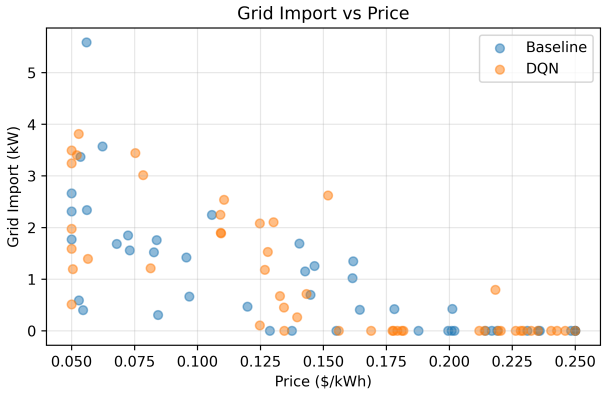

Fig. 9. Sensitivity of grid import to electricity price

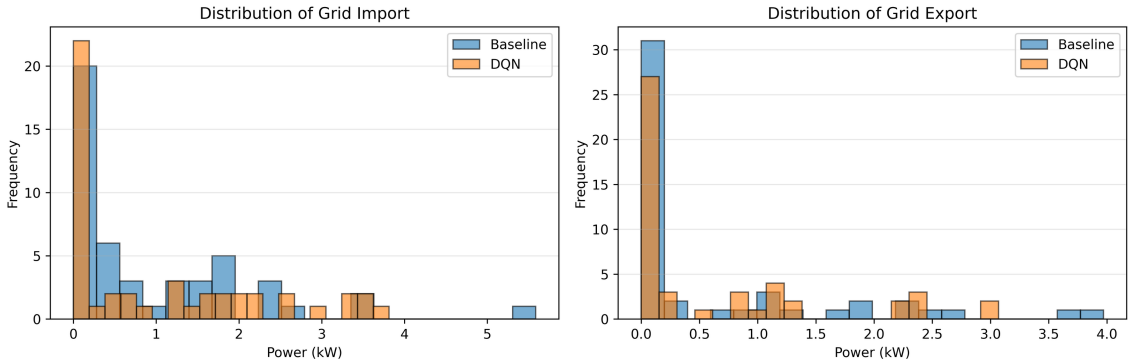

Fig. 10. Grid import/export power distributions

The concentration of imports near zero at mid prices arises from the preference for storing energy during low price hours. The plots confirm the dispatch logic observed in the SoC trajectory.

Table 4 shows that the DQN policy reduces peak grid current by 31.7% (24.280 A → 16.580 A) while slightly increasing the average grid current (6.691 A → 7.160 A). Fig. 11 corroborates this peak-shaving behaviour: import spikes are clipped during high-price periods, shifting power away from the most stressed intervals. On the storage side, Table 4 indicates a marked rise in battery throughput (average battery current 26.042 A → 55.990 A) with identical hardware-limited peaks (62.5 A) in both policies, consistent with Fig. 11's long discharge plateaus. The more negative average net battery current (−54.688 A vs −1.302 A) confirms net discharge by DQN over the horizon.

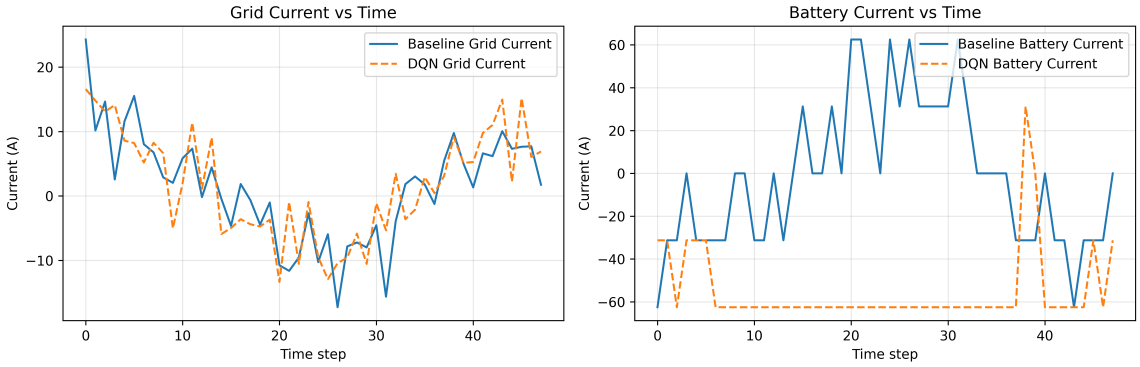

Fig. 11. Currents over time (grid & battery)

Table 4. Current envelope and throughput (grid & battery)

Metric	Baseline	DQN
Average grid current (A)	6.690936	7.160372
Maximum grid current (A)	24.280140	16.579518
Average battery current (A)	26.041667	55.989583
Maximum battery current (A)	62.500000	62.500000
Average net grid current (A)	1.368340	1.784345

The lower peak grid current improves thermal margins of cables and converters because peak heating is reduced when the highest current episodes are clipped. The increase in average battery current reflects intentional scheduling near high value events. The pattern is consistent with targeted use of stored energy rather than high frequency cycling.

Although the mean frequency remains essentially unchanged (49.994 vs 49.992 Hz), Table 5 shows that DQN improves the nadir (minimum frequency) from 49.888 Hz to 49.924 Hz and reduces the zenith from 50.079 Hz to 50.061 Hz. Expressed as absolute deviations from 50 Hz, the nadir error drops by ≈31.7%, and the zenith error by ≈22.7%. These gains align with the current-peak suppression seen in Table 6 and Fig. 11, and they are visible as tighter excursions in Fig. 12. The DQN therefore attenuates power-imbalance–driven frequency departures without altering the steady-state setpoint. The reduction in frequency deviation enlarges the safe region around the nominal operating point. The gain represents a mitigated response to power imbalance because the dispatch logic avoids abrupt grid power movements.

Table 5. Frequency stability metrics (mean, nadir, zenith)

Metric	Baseline	DQN
Average frequency (Hz)	49.993706	49.991792
Minimum frequency (Hz)	49.888311	49.923734
Maximum frequency (Hz)	50.079407	50.061362

Table 6 reports a modest rise in the mean phase angle (49.16° → 49.53°) alongside a narrower operating range: the minimum increases substantially (5.36° → 22.69°) while the maximum moves to the limit (84.70° → 90.00°). The net effect, also apparent in Fig. 12, is a ~15% contraction of the

phase-angle spread, which reduces angular stress and enhances synchronism margins at the PCC. Together with Table 5, these data indicate that the DQN policy yields better small-signal quality (frequency/angle excursions) by reallocating power dynamically.

Table 6. Voltage phase-angle statistics (mean, min, max)

Metric	Baseline	DQN
Average phase angle (deg)	49.157586	49.525423
Minimum phase angle (deg)	5.363559	22.688572
Maximum phase angle (deg)	84.695442	90.000000

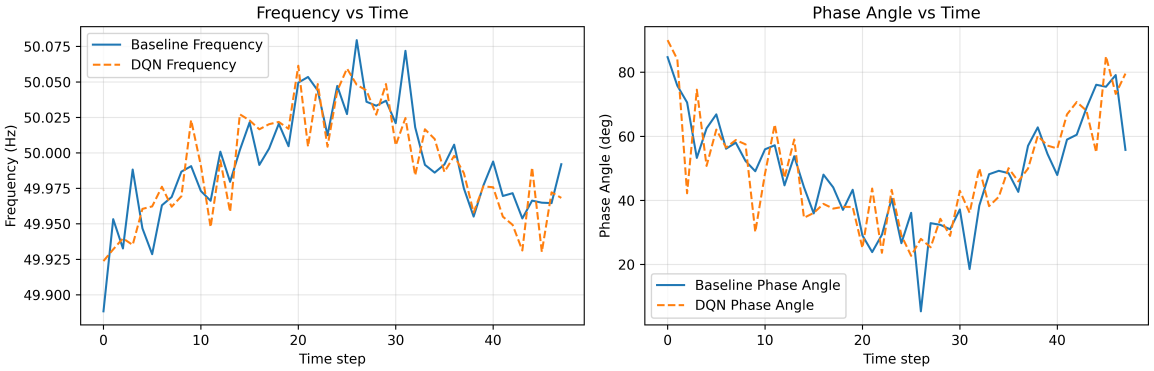

Fig. 12. Currents over time (grid & battery)

The study reports current frequency and phase angle responses together with economic performance. These indicators seldom appear together in previous reinforcement learning based energy management works. The combination forms a fuller view of the controller behavior. The main findings concern the shift of battery usage into targeted intervals that carry high value within the daily cycle. The shift improves cost and reduces peak grid current. The comparison with existing studies shows that prior work often evaluates economic indicators only. The present analysis expands the set of indicators by quantifying electrical quality metrics at the same operating point. The implication is a dispatch strategy that improves reliability because large excursions in grid current and frequency become less prominent. The limitation lies in the absence of detailed battery ageing cost and the use of a single synthetic profile. The extension to wider datasets and the addition of degradation aware constraints will enrich the analysis in future investigations.

4. Conclusion

This study evaluates a Deep Q Network energy management policy for a grid connected photovoltaic microgrid and shows that the learned controller can reduce the magnitude of daily operating cost by about 35.6 % relative to a transparent rule-based baseline while keeping the state of charge within prescribed bounds. The policy shifts charging toward low price periods and schedules discharge near expensive intervals so the microgrid imports more energy when electricity is cheap and relies on the battery when the grid is costly. The results also indicate that the agent attenuates current peaks at the point of common coupling and tightens the range of frequency and phase angle excursions which supports a more benign operating regime for grid components and converters. These findings suggest that a carefully engineered discrete action DQN can coordinate economic objectives and electrical quality metrics in a single decision process that remains compatible with engineering practice.

The work contributes a compact Markov decision process model for a single bus microgrid that embeds battery degradation penalties branch current envelopes frequency deviation and phase angle variation in the running cost so that the agent optimizes a multi criteria objective rather than a purely monetary one. The numerical study provides a joint view of cost state of charge trajectories current

statistics and small signal frequency behaviour under a common experimental protocol that uses the same traces seeds and constraints for the baseline and the DQN policy. The implementation uses NumPy arrays and explicit update equations so every state transition action selection and cost evaluation can be inspected statement by statement which is useful for debugging auditing and porting the controller to embedded platforms that do not host large machine learning runtimes. This design reduces dependence on opaque computation graphs and eases translation of the algorithm to C or other low-level languages that are common in industrial controllers.

The study has several limitations that define a clear path for future research. The environment relies on synthetic photovoltaic load and price profiles for a single day horizon so the reported improvements may change under different climates tariff structures or longer horizons and the battery model uses a simplified degradation surrogate that does not capture the full physics of aging. The learning algorithm remains a basic Deep Q Network without Double DQN dueling heads or prioritized replay so stability and sample efficiency may be improved further. Future work will investigate richer battery health models and more advanced value learning schemes apply the framework to measured microgrid data sets and explore deployment on hardware in the loop or field demonstrators. Extensions to multi agent settings and to larger hybrid alternating current and direct current architectures can generalize the proposed approach and support cooperative scheduling policies that address both operational efficiency and long-term reliability in real energy systems.

Author Contribution: The author conceived, conducted, and wrote the entire paper.

Funding: This research received no external funding.

Acknowledgment: The author gratefully acknowledges the support of the Thai Nguyen University of Technology and the Education Technology and Adaptive Learning Institute, Thai Nguyen University of Technology, Vietnam.

Conflicts of Interest: The author declare no conflict of interest.

References

[1] G. Li and A. Z. Ahmad, "Energy management system for grid-connected microgrids with deep reinforcement learning," in *2025 International Conference on Intelligent Transportation and New Energy Technology (ITNET)*, Nanning, China, 2025, pp. 201–206, https://doi.org/10.1109/ITNET65199.2025.11162769.

[2] M. R. M. Altimania, A. Basem, B. Saydullaev *et al*., "Adaptive multi-objective optimization of microgrid energy management using deep reinforcement learning considering battery degradation and renewable uncertainty," *Research Square Preprint*, Jun. 2025, https://doi.org/10.21203/rs.3.rs-6744762/v1.

[3] O. A. Talab and İ. Avci, "Energy management in microgrids using model-free deep reinforcement learning approach," *IEEE Access*, vol. 13, pp. 5871–5891, 2025, https://doi.org/10.1109/ACCESS.2025.3525843.

[4] J. Li, Z. Jiang, Z. Chen, J. Liu, and L. Cheng, "CuEMS: Deep reinforcement learning for community control of energy management systems in microgrids," *Energy and Buildings*, vol. 304, p. 113865, Feb. 2024, https://doi.org/10.1016/j.enbuild.2023.113865.

[5] Y. Pei, Y. Yao, J. Zhao, F. Ding, and J. Wang, "Deep reinforcement learning for microgrid cost optimization considering load flexibility," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*, Seattle, WA, USA, Jul. 2024, pp. 1–5, https://doi.org/10.1109/PESGM51994.2024.10688837.

[6] A. Saxena, S. Aswini, S. V. Singh, A. K. Aravinda, M. Gupta, and M. A. Alkhafaji, "Utilizing deep reinforcement learning for energy trading in microgrids," in *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*, Pune, India, Mar. 2024, pp. 1–6, https://doi.org/10.1109/TQCEBT59414.2024.10545306.

[7] C. P. Agupugo, M. F. C. Tochukwu, K. A. Ogunmoye, A. S. Mosha, and F. Sabbih, "Review of smart microgrid platform integrating AI and deep reinforcement learning for sustainable energy management,"

International Journal of Future Engineering Innovations, vol. 2, no. 3, pp. 1–17, 2025, https://doi.org/10.54660/IJFEI.2025.2.3.01-17.

[8] Y. Zheng, J. Jia, and D. An, "Energy management for microgrids with hybrid hydrogen-battery storage: A reinforcement learning framework integrated multi-objective dynamic regulation," *Processes*, vol. 13, no. 8, p. 2558, Aug. 2025, https://doi.org/10.3390/pr13082558.

[9] A. Yang, Z. Lin, K. Lin, and L. Li, "Optimal energy scheduling of a microgrid based on offline-to-online deep reinforcement learning," in *2024 6th International Conference on Energy Systems and Electrical Power (ICESEP)*, Wuhan, China, Jun. 2024, pp. 1088–1092, https://doi.org/10.1109/ICESEP62218.2024.10651738.

[10] J. Si, F. Deng, Z. Deng, H. Li, and F. Liu, "A microgrid energy storage technology based on reinforcement learning," in *2024 6th International Conference on Energy, Power and Grid (ICEPG)*, Guangzhou, China, Sep. 2024, pp. 370–373, https://doi.org/10.1109/ICEPG63230.2024.10775475.

[11] A. T. Hassan, F. A. Banakhr, M. M. Mahmoud, M. I. Mosaad, A. F. Rashwan, M. R. Mosa, M. M. Hussein, and T. H. Mohamed, "Adaptive load frequency control in microgrids considering PV sources and EV impacts: Applications of hybrid sine cosine optimizer and balloon effect identifier algorithms," *International Journal of Robotics and Control Systems*, vol. 4, no. 2, pp. 941–957, 2024, https://doi.org/10.31763/ijrcs.v4i2.1448.

[12] M. N. A. Hamid, F. A. Banakhr, T. H. Mohamed, S. M. Ali, M. M. Mahmoud, M. I. Mosaad, A. A. H. Albla, and M. M. Hussein, "Adaptive frequency control of isolated microgrids implementing different recent optimization techniques," *International Journal of Robotics and Control Systems*, vol. 4, no. 3, pp. 1000–1012, 2024, https://doi.org/10.31763/ijrcs.v4i3.1432.

[13] M. Zadehbagheri, A. Ma'arif, M. J. Kiani, and A. A. Poorat, "Adaptive droop control strategy for load sharing in hybrid microgrids," *International Journal of Robotics and Control Systems*, vol. 3, no. 1, pp. 74–83, 2023, https://doi.org/10.31763/ijrcs.v3i1.838.

[14] T. Saha, A. Haque, M. A. Halim, and M. M. Hossain, "A review on energy management of community microgrids using adaptable renewable energy sources," *International Journal of Robotics and Control Systems*, vol. 3, no. 4, pp. 824–838, 2023, https://doi.org/10.31763/ijrcs.v3i4.1009.

[15] M. S. Akter, A. M. Islam, and M. M. Hasan, "Microgrid energy management using weather forecasts: Case study, discussion, and challenges," *International Journal of Robotics and Control Systems*, vol. 3, no. 4, pp. 749–764, 2023, https://doi.org/10.31763/ijrcs.v3i4.1000.

[16] K. B. Kwon and H. Zhu, "Reinforcement learning-based optimal battery control under cycle-based degradation cost," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4909–4917, Nov. 2022, https://doi.org/10.1109/TSG.2022.3180674.

[17] H. Xiong, J. Chen, S. Rong, and A. Zhang, "Power battery scheduling optimization based on double DQN algorithm with constraints," *Applied Sciences*, vol. 13, no. 13, p. 7702, Jun. 2023, https://doi.org/10.3390/app13137702.

[18] D. K. Panda, O. Turner, S. Das, and M. Abusara, "Prioritized experience replay-based deep distributional reinforcement learning for battery operation in microgrids," *Journal of Cleaner Production*, vol. 434, p. 139947, Nov. 2023, https://doi.org/10.1016/j.jclepro.2023.139947.

[19] M. Sage and Y. F. Zhao, "Deep reinforcement learning for economic battery dispatch: A comprehensive comparison of algorithms and experiment design choices," *Journal of Energy Storage*, vol. 115, p. 115428, 2025, https://doi.org/10.1016/j.est.2025.115428.

[20] A. Farhana, N. Satheesh, M. Ramya, J. V. N. Ramesh, and Y. A. B. El Ebiary, "Efficient deep reinforcement learning for smart buildings: Integrating energy storage systems through advanced energy management strategies," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 12, pp. 548–559, 2023, https://doi.org/10.14569/IJACSA.2023.0141257.

[21] M. M. Rahman, M. M. Hasan, Y. Suleymanov, S. Dadon, S. Saha, and T. T. Suki, "Energy purchase optimization for microgrid systems using deep Q-learning," in *Proceedings of the 2025 International Conference on Clean Electrical Power (ICCEP)*, Villasimius, Italy, Jun. 2025, pp. 1–5, https://doi.org/10.1109/ICCEP65222.2025.11143685.

[22] A. Selim, H. Mo, H. Pota, and D. Dong, "Optimal scheduling of battery energy storage systems using a reinforcement learning-based approach," *IFAC PapersOnLine*, vol. 56, no. 2, pp. 11741–11747, 2023, https://doi.org/10.1016/j.ifacol.2023.10.546.

[23] N. Basil, B. M. Sabbar, H. M. Marhoon, A. F. Mohammed, and A. Ma'arif, "Systematic review of unmanned aerial vehicles control: Challenges, solutions, and meta-heuristic optimization," *International Journal of Robotics and Control Systems*, vol. 4, no. 4, pp. 1794–1818, Oct. 2024, https://doi.org/10.31763/ijrcs.v4i4.1596.

[24] M. A. Mossa, O. Gam, and N. Bianchi, "Performance enhancement of a hybrid renewable energy system accompanied by an energy storage unit using an effective control system," *International Journal of Robotics and Control Systems*, vol. 2, no. 1, pp. 140–171, Feb. 2022, https://doi.org/10.31763/ijrcs.v2i1.599.

[25] S. Ekinci, E. Eker, D. Izci, A. Smerat, and L. Abualigah, "Enhanced RSA-optimized TID controller for frequency stabilization in a two-area power system," *International Journal of Robotics and Control Systems*, vol. 4, no. 4, pp. 1886–1902, Nov. 2024, https://doi.org/10.31763/ijrcs.v4i4.1644.

[26] M. I. M. Ameerdin, M. H. Jamaluddin, A. Z. Shukor, and S. Mohamad, "A review of deep learning-based defect detection and panel localization for photovoltaic panel surveillance systems," *International Journal of Robotics and Control Systems*, vol. 4, no. 4, pp. 1746–1771, Oct. 2024, https://doi.org/10.31763/ijrcs.v4i4.1579.

[27] R. Moumni, K. Laroussi, I. Benlaloui, M. M. Mahmoud, and M. F. Elnaggar, "Optimizing single-inverter electric differential systems for electric vehicle propulsion applications," *International Journal of Robotics and Control Systems*, vol. 4, no. 4, pp. 1772–1793, Oct. 2024, https://doi.org/10.31763/ijrcs.v4i4.1542.

[28] O. Akbulut, M. Çavuş, M. Cengiz, A. Allahham, D. Giaouris, and M. Forshaw, "Hybrid intelligent control system for adaptive microgrid optimization: Integration of rule-based control and deep learning techniques," *Energies*, vol. 17, no. 10, p. 2260, May 2024, https://doi.org/10.3390/en17102260.

[29] P. P. Kumar, R. S. S. Nuvvula, Sk. A. Shezan, S. R. Ahammed, B. J. M., V. Satyanarayana, and A. Ali, "Intelligent energy management system for microgrids using reinforcement learning," in *Proceedings of the 12th International Conference on Smart Grid (icSmartGrid)*, Setúbal, Portugal, May. 2024, pp. 322–328, https://doi.org/10.1109/icSmartGrid61824.2024.10578215.

[30] P. A. Babu and R. Ayyasamy, "Power control and optimization for power loss reduction using deep learning in microgrid systems," *Electric Power Components and Systems*, vol. 52, no. 2, pp. 219–232, Jul. 2023, https://doi.org/10.1080/15325008.2023.2217175.

[31] B. Zhang, Z. Chen, and A. M. Y. Ghias, "Deep reinforcement learning-based energy management strategy for a microgrid with flexible loads," in *Proceedings of the 2023 International Conference on Power Energy Systems and Applications (ICoPESA)*, Nanjing, China, Feb. 2023, pp. 187–191, https://doi.org/10.1109/ICoPESA56898.2023.10141490.

[32] M. W. ul Hassan, M. F. Farhan, Z. Ahmed, T. Abid, M. A. Iqbal, and M. S. Ashraf, "Deep reinforcement learning for control of microgrids: A review," *Lahore Garrison University Research Journal of Computer Science and Information Technology*, vol. 6, no. 4, pp. 45–61, Dec. 2022, https://doi.org/10.54692/lgurjcsit.2022.0604359.

[33] A. A. Ladjici and A. Tiguercha, "Deep reinforcement learning for microgrid power management systems," in *Proceedings of the 9th International Conference on Control, Decision and Information Technologies (CoDIT)*, Rome, Italy, Jul. 2023, pp. 1144–1149, https://doi.org/10.1109/CoDIT58514.2023.10284334.

[34] F. Yao, W. Zhao, M. Forshaw, and Y. Song, "A holistic power optimization approach for microgrid control based on deep reinforcement learning," *arXiv preprint*, arXiv:2403.01013, Mar. 2024, https://doi.org/10.48550/arXiv.2403.01013.

[35] M. F. Dong, J. Li, W. Huang, W. Y. Lin, J. C. Liao, and C. Y. Hsueh, "Optimal economic energy management of microgrids using deep deterministic policy gradient," in *Proceedings of the 7th*

International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, Jun. 2025, pp. 1–4, https://doi.org/10.1109/IS3C65361.2025.11130992.

[36] S. W. Jung, Y. Y. An, B. Suh, Y. B. Park, J. Kim, and K. I. Kim, "Multi-agent deep reinforcement learning for scheduling of energy storage systems in microgrids," *Mathematics*, vol. 13, no. 12, p. 1999, Jun. 2025, https://doi.org/10.3390/math13121999.

[37] X. Zhou, J. Wang, X. Wang, and S. Chen, "Deep reinforcement learning for microgrid operation optimization: A review," in *Proceedings of the 8th Asia Conference on Power and Electrical Engineering (ACPEE)*, Chongqing, China, Apr. 2023, pp. 2059–2065, https://doi.org/10.1109/ACPEE56931.2023.10135713.

[38] N. F. P. Dinata, M. I. Jambak, M. A. M. Ramli, and M. A. B. Sidik, "Utilizing deep reinforcement learning for enhanced microgrid voltage regulation under fluctuating load conditions," in *Proceedings of the 2024 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Palembang, Indonesia, Sept. 2024, pp. 343–348, https://doi.org/10.1109/ICECOS63900.2024.10791118.

[39] I. Ahmed, A. Pedersen, and L. Mihet-Popa, "Smart microgrid optimization using deep reinforcement learning by utilizing energy storage systems," in *2024 4th International Conference on Smart Grid and Renewable Energy (SGRE)*, 2024, pp. 1–7, https://doi.org/10.1109/SGRE59715.2024.10428874.

[40] S. Ramesh, B. N. Sukanth, S. J. Sathyavarapu, V. Sharma, A. A. N. Kumaar, and M. Khanna, "Comparative analysis of Q-learning, SARSA, and deep Q-network for microgrid energy management," *Scientific Reports*, vol. 15, p. 83625, Jan. 2025, https://doi.org/10.1038/s41598-024-83625-8.

[41] J. Qi, L. Lei, K. Zheng, and S. X. Yang, "Joint energy dispatch and unit commitment in microgrids based on deep reinforcement learning," *arXiv preprint*, arXiv:2206.01663, 2023, https://doi.org/10.48550/arXiv.2206.01663.

[42] A. Rizki, A. Touil, A. Echchatbi, and R. Oucheikh, "A reinforcement learning approach based on group relative policy optimization for economic dispatch in smart grids," *Electricity*, vol. 6, no. 3, p. 49, Sept. 2025, https://doi.org/10.3390/electricity6030049.

[43] K. Singh, A. Nagar, G. Jindal, and N. K. Saraswat, "Artificial intelligence-based methods for economic power dispatch in smart grids," in *Proceedings of the 7th International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Sonepat, India, Apr. 2025, pp. 33–39, https://doi.org/10.1109/CCICT65753.2025.00016.

[44] P. Dai, W. Yu, G. Wen, and S. Baldi, "Distributed reinforcement learning algorithm for dynamic economic dispatch with unknown generation cost functions," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2258–2267, Apr. 2020, https://doi.org/10.1109/TII.2019.2933443.

[45] Y. Fu, X. Guo, Y. Mi, M. Yuan, X. Ge, X. Su, and Z. Li, "The distributed economic dispatch of smart grids based on deep reinforcement learning," *IET Generation, Transmission & Distribution*, vol. 15, no. 18, pp. 2645–2658, Sept. 2021, https://doi.org/10.1049/GTD2.12206.

[46] Q. Xu, C. Yu, X. Yuan, Z. Fu, and H. Liu, "Distributed Q-learning algorithm for economic dispatch of smart grids with unknown cost functions," in *Proceedings of the 2022 China Automation Congress (CAC)*, Xiamen, China, Nov. 2022, pp. 6493–6497, https://doi.org/10.1109/CAC57257.2022.10055962.

[47] G. Wen, X. Yu, P. Dai, and W. Yu, "Economic dispatch of smart grids with unknown cost functions and switching network topology," in *Proceedings of the 4th International Conference on Data Driven Optimization of Complex Systems (DOCS)*, Guilin, China, Oct. 2022, pp. 1–6, https://doi.org/10.1109/DOCS55193.2022.9967783.

[48] Z. Yu, G. Zhang, T. Xiao, X. Wang, and H. Zhong, "Dynamic economic dispatch considering demand response based on reinforcement learning," in *Proceedings of the 2021 International Conference on Power System Technology (POWERCON)*, Kunming, China, Dec. 2021, pp. 1941–1947, https://doi.org/10.1109/POWERCON53785.2021.9697597.

[49] S. Santosh, "Energy optimization for smart grids using reinforcement learning," in *Proceedings of the 6th International Conference for Emerging Technology (INCET)*, Hubli, India, May. 2025, pp. 1–5, https://doi.org/10.1109/INCET64471.2025.11140071.

[50] A. H. Henni, B. Boukezata, J. Gaber, K. Henni, and P. Lorenz, "Optimization of energy costs using deep reinforcement learning in smart grids," in *Proceedings of the World Conference on Complex Systems (WCCS)*, Oum El Bouaghi, Algeria, Nov. 2024, pp. 1–6, https://doi.org/10.1109/WCCS62745.2024.10765521.

[51] L. Ding, Z. Lin, and G. Yan, "Multi-agent deep reinforcement learning algorithm for distributed economic dispatch in smart grids," in *Proceedings of the 46th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Singapore, Oct. 2020, pp. 3529–3534, https://doi.org/10.1109/IECON43393.2020.9255238.

[52] C. Hu, G. Wen, S. Wang, J. Fu, and W. Yu, "Distributed multi-agent reinforcement learning with action networks for dynamic economic dispatch," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 9553–9564, Jul. 2023, https://doi.org/10.1109/TNNLS.2023.3234049.

[53] G. C. Liao, B. T. Liao, Z. Q. Liang, A. T. Yeh, and R. C. Wu, "Applying deep reinforcement learning to solve the low-carbon economic dispatch problem in smart microgrids," *Journal of Physics: Conference Series*, vol. 2878, p. 012005, Oct. 2024, https://doi.org/10.1088/1742-6596/2878/1/012005.

[54] F. Yang, G. Xiaoyan, M. Yang, X. Zhang, Y. Mi, Y. Wang, X. Chen, and L. Yu, "Fully distributed intelligent power grid economic dispatching method based on deep reinforcement learning," *Chinese Patent,* CN111236953A, Mar. 2020, https://eureka.patsnap.com/patent-CN110929948A.

[55] Z. Yin, M. Wang, Y. Lv, L. Li, J. Ren, and H. Cao, "An online economic scheduling method based on a Lagrangian relaxation reinforcement algorithm," in *Proceedings of the 2024 International Conference on Power System Technology (ICPST)*, Wuhan, China, May. 2024, pp. 2144–2151, https://doi.org/10.1109/ICPST61417.2024.10602241.

[56] J. Qin, H. Liu, H. Meng, W. Gu, Q. Xu, and W. Yu, "Robust dynamic economic dispatch in smart grids using intelligent learning technology," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 4, pp. 3759–3770, 2024, https://doi.org/10.1109/TNSE.2024.3384505.

[57] B. Liu, "Deep reinforcement learning for intelligent load balancing in smart power grids," *IEEE Access*, vol. 13, pp. 164170–164185, 2025, https://doi.org/10.1109/ACCESS.2025.3606914.

[58] F. Sangoleye, J. Jao, K. A. Faris, E. E. Tsiropoulou, and S. Papavassiliou, "Reinforcement learning-based demand response management in smart grid systems with prosumers," *IEEE Systems Journal*, vol. 17, no. 2, pp. 1797–1807, Jun. 2023, https://doi.org/10.1109/JSYST.2023.3248320.

[59] D. Tabas and B. Zhang, "Computationally efficient safe reinforcement learning for power systems," in *Proceedings of the 2022 American Control Conference (ACC)*, Atlanta, GA, USA, Jun. 2022, pp. 3303–3310, https://doi.org/10.23919/ACC53348.2022.9867652.

[60] A. R. M. Matavalam, K. P. Guddanti, Y. Weng, and V. Ajjarapu, "Curriculum-based reinforcement learning of grid topology controllers to prevent thermal cascading," *IEEE Transactions on Power Systems*, vol. 38, no. 5, pp. 4206–4220, Sept. 2023, https://doi.org/10.1109/TPWRS.2022.3213487.

[61] P. Yu, Z. Wang, H. Zhang, and Y. Song, "Safe reinforcement learning for power system control: A review," *arXiv preprint*, arXiv:2407.00681, 2024, https://doi.org/10.48550/arXiv.2407.00681.

[62] Y. Zhou, L. Zhou, D. Shi, and X. Zhao, "Coordinated frequency control through safe reinforcement learning," in *Proceedings of the IEEE Power & Energy Society General Meeting (PESGM)*, Denver, CO, USA, Jul. 2022, pp. 1–5, https://doi.org/10.1109/PESGM48719.2022.9916894.

[63] A. Dwivedi, S. Paternain, and A. Tajer, "Blackout mitigation via physics-guided reinforcement learning," *IEEE Transactions on Power Systems*, vol. 40, no. 3, pp. 2363–2375, 2025, https://doi.org/10.1109/TPWRS.2024.3472570.

[64] M. Cauz, A. Bolland, N. Wyrsch, and C. Ballif, "Reinforcement learning for efficient design and control co-optimisation of energy systems," *arXiv preprint*, arXiv:2406.19825, 2024, https://doi.org/10.48550/arXiv.2406.19825.

[65] M. Gautam, "Deep reinforcement learning for resilient power and energy systems: Progress, prospects, and future avenues," *Electricity*, vol. 4, no. 4, pp. 336–380, Dec. 2023, https://doi.org/10.3390/electricity4040020.