Data-Driven Autonomous Voltage Control in Off-Grid PV Microgrids Using Soft Actor-Critic Reinforcement Learning

Nayana Shetty
NITTE (Demeed to be University)
NMAM Institute of Technology , Department of Electrical & Electronics Engineering
Mangalore, India,
nayanashetty@nitte.edu.in

Nandini S Patil
Department of Electrical and Electronic Engineering Technology
Ballari Institute of Technology and Management
nandiniparu@gmail.com

Ahmed Ali
Department of Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, SOUTH AFRICA
aali@uj.ac.za

***Abstract*—The conventional droop controllers are poor for the control problem of the off-grid PV-battery microgrids, that is: keeping bus voltage within safe limits and at the same time fully utilized solar generation. Tight voltage bounds cause conservative battery dispatch, which increases unmet load from renewables. This paper presents a framework for deep reinforcement learning with continuous battery power control in standalone PV-battery microgrid systems, comparing results from the Soft Actor-Critic (SAC) algorithm against a Proximal Policy Optimization (PPO) baseline and an existing rule-based droop controller. We formulate the control environment as a Markov decision process with an A six-dimensional state — normalised PV generation, aggregate load, battery SOC, bus voltage and sinusoidal time encodings. The Terminal Voltage is expected to be linked with SOC through a model of linear open-circuit voltage so that the agent can then learn how to control SOC proactively and avoid overvoltage instead of reactively compensating for it. The framework is applied to a single year of actual operational data from a three-household isolated microgrid on Réunion Island — 7,710 hourly records, 5 held out test episodes. SAC attained a 100% RER with no grid import and zero voltage violations for all the test episodes — which is improvement by 7.74 percentage points over the droop controller (92.26% RER). It also found its first episode of 100% RER without violating either rule at training episode 3. In episode 4, PPO only recorded 10 violations and never exceeded an RER of 88.47% during training: the entropy-regularized off-policy learning clearly works in their favor with rare, early exploration events.**

I. INTRODUCTION

The battery is everything, especially in remote island communities. It stores solar excess, it can support bus voltage and supply load when generation declines — all at the same time while having no grid to fall back on [1], [2]. This is the crux of control: these objectives are in conflict. When morning PV increases faster than load or net demand, the battery charge aggressively to absorb excess — but too aggressive and terminal voltage rises above thresholds [3], risking electrode sulfation for lead-acid systems and accelerated degradation [4]. If the controller backs off to save voltage, it loses generation without supply to serve the load. Conventional droop controllers deal with this limited foresight through a simple proportional rule — charge harder when SOC is low, ease off when it is high — but have zero knowledge of upcoming hours passing [5]. This leads to an inherent trade-off: higher voltage margins require a lower renewable energy ratio (RER) which cannot be resolved by tuning the droop gain [6].

An alternate strategy of deep reinforcement learning provides an agent to learn about the intra-day generation cycle and control SOC proactively, rather than reactively [7], [8]. At a high level, Microgrid scheduling has been effectively solved through PPO and DQN variants with rather significant improvements over rule based baselines [9], [10]. However as far as voltage control is concerned, three remain: most DRL studies consider grid-connected systems where voltage excursions simply export to the grid [11], on-policy methods are slow to converge when constraint violations are rare events such as the semi-analytical approach of [12] and discretised action spaces lead to quantisation errors that interfere with fine-grained voltage management [13]. Autonomous off-grid deployment needs are still well ahead of published DRL results.

This paper proposes an SAC-based voltage control framework for isolated PV-battery microgrids [14]. SAC's entropy regularisation promotes exploratory diversity during early training — precisely when the agent needs to encounter and recover from rare overvoltage events — while its off-policy replay buffer makes much more efficient use of those rare transitions than PPO's on-policy updates [15]. The agent controls continuous battery power with SOC-linked voltage dynamics, validated on one full year of real operational data from a three-household microgrid on Réunion Island [16]. Three contributions follow: (i) formulation of off-grid voltage regulation as a continuous-action MDP with SOC-based voltage dynamics; (ii) empirical demonstration that SAC converges to a violation-free 100% RER policy by episode 3 — versus PPO's 10 violations in episode 4 and peak training RER of 88.47%; and (iii) a reproducible three-way benchmark (SAC vs PPO vs droop) on real open-access microgrid data [17].

II. METHODOLOGY

2.1 Deep RL for Microgrid Energy Management

Early DRL microgrid work used discretised DQN variants, cutting costs 15–30% over rule-based schedulers in simulation [18]. More recent actor-critic approaches handle continuous dispatch variables — TD3 agents have improved energy trading profits in grid-connected PV-wind-battery systems [19], and A2C agents have generalised across seasonal load variations without forecasting modules [20]. The shared limitation is that almost all of this work assumes grid connection. When voltage excursions happen, they transfer to the utility network rather than being resolved locally. That assumption breaks completely in islanded off-grid systems where the battery controller is the only voltage-regulating element.

2.2 RL for Voltage Regulation in Isolated Microgrids

Voltage control in an islanded microgrid is fundamentally harder than cost minimisation — violations have immediate physical consequences, not abstract financial ones. Droop controllers are the industrial standard but they react to current SOC without any awareness of upcoming generation ramps [8]. MPC addresses this through forecasting but needs accurate prediction models that rarely exist in remote deployments [21]. Among learning-based approaches, data-driven decentralised voltage control has shown faster regulation than model-based methods [22], and multi-agent DRL has improved disturbance recovery in meshed AC microgrids [23]. Both assume either grid support or full inter-agent communication — neither applies to a single-battery islanded system. Crucially, all on-policy methods (PPO, A3C) surveyed for voltage regulation reported slow convergence in sparse-violation settings, needing 50–200 episodes before producing consistently safe trajectories.

2.3 SAC for Energy Systems and Gap

SAC has outperformed PPO and TD3 in sparse-reward energy applications — HVAC demand response [24], EV charging [25], and building optimisation [26] — consistently benefiting from entropy regularisation when the optimal policy mixes conservative and aggressive actions. SAC-trained HVAC controllers maintained thermal comfort 98% of the time while cutting energy use by 21% versus PPO [24]. Despite this, no published study has applied SAC to hard voltage constraint management in an off-grid PV-battery microgrid with real annual operational data. Existing SAC energy work uses soft cost or comfort constraints — not the hard voltage bounds that must hold at every timestep to prevent hardware damage. No study benchmarks SAC against both a droop controller and a PPO agent on identical real microgrid episodes [27], [28]. This paper closes all three gaps simultaneously.

2.4 System Overview and Problem Formulation

The Mafate off-grid microgrid serves three residential households through two parallel PV arrays and a lead-acid battery bank on a common DC bus — no grid connection. The battery is the only controllable element: its charge/discharge rate determines bus voltage through the terminal voltage relationship, and its SOC trajectory determines whether enough energy is available overnight.

The control problem can be understood as a finite-horizon MDP over hourly timesteps—each episode represents one 24 hour day and has 24 decision steps. The agent at each step sees a state, chooses a continuous battery power command, gets a reward and goes to the next state. An explicit module for forecasting is not required — the agent learns intra-day generation patterns from time encodings and replay buffer experience.

III. System Modeling and Formulation

3.2 Environment Design

State: Six-dimensional vector at each hourly step t:

$$\mathbf{s}_t = \left[\frac{P_{pv,t}}{P_{pv,max}},\ \frac{P_{load,t}}{P_{load,max}},\ \mathrm{SOC}_t,\ \frac{V_t - V_{min}}{V_{max} - V_{min}},\ \sin\left(\frac{2\pi h_t}{24}\right),\ \cos\left(\frac{2\pi h_t}{24}\right)\right] \quad (1)$$

Action: Continuous scalar $a_t \in [-1, +1]$ maps to physical battery power $P_{bat,t} = a_t \times 3.0$ kW. Positive = charging, negative = discharging.

SOC dynamics:

$$\mathrm{SOC}_{t+1} = \mathrm{SOC}_t + \frac{P_{bat,t} \cdot \Delta t}{E_{bat}} \quad (2)$$

Clipped to [0.10, 0.90]. Bus voltage follows a linear open-circuit model: $V_t = 49.3 + \mathrm{SOC}_t \cdot (62.01 - 49.3)$ V.

Reward: Combines renewable utilisation, voltage violation penalty, and grid import cost:

$r_t = 1.0 \cdot \mathrm{RER}_t - 5.0 \cdot \{V_t \notin [50.0, 60.0]\} - 2.0 \cdot \mathrm{Pimport}_t$

Violation avoidance ($w_2 = 5.0$) is weighted highest to prioritise voltage safety.

3.3 Soft Actor-Critic

SAC maximises an entropy-augmented objective:

$$\pi^* = \underset{\pi}{\mathrm{argmax}} \mathbb{E}_\pi \left[\sum_t \gamma^t \left(r_t + \alpha \mathcal{H}\left(\pi(\cdot\,|\mathbf{s}_t)\right)\right)\right] \quad (3)$$

where $\gamma = 0.99$ and $\alpha = 0.20$. The entropy term keeps the policy exploratory during early training — critical when overvoltage events are rare and the agent needs to learn from them rather than avoid them entirely. Three networks: stochastic actor (two hidden layers, 256 units, ReLU, tanh output) and two independent critics using clipped double-Q to reduce overestimation. Replay buffer capacity 10,000, mini-batch 64, Adam lr = 5×10^{-4}, soft target update $\tau = 0.005$, five gradient steps per environment step.

3.4 PPO Baseline

PPO uses the clipped surrogate objective with probability ratio $\rho_t = \pi\theta(a_t|s_t)/\pi\theta\mathrm{old}(a_t|s_t)$:

$$\mathcal{L}^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right] \quad (4)$$

GAE with $\lambda = 0.95$, entropy coefficient 0.01, 10 update epochs per cycle, mini-batch 64, two-layer MLP backbone of 64 units shared between actor and critic.

3.5 Rule-Based Droop Controller

Battery power is proportional to SOC deviation from a 0.5 setpoint:

$P_{\beta at}$,droop $= 2.0 \cdot (\mathrm{SOC}_t - 0.5) \cdot 3.0$ kW

Simple and fast but entirely reactive — no anticipation of upcoming generation ramps or load patterns.

3.6 Data and Training Protocol

One-minute Mafate microgrid data (May 2020–May 2021) is resampled to hourly means. PV output sums both arrays; load aggregates all three households. After removing missing-value records, 7,710 valid hourly observations remain, split into 321 complete 24-hour episodes — 306 for training (60 training epochs), 15 held out for testing. All three controllers are evaluated on identical held-out episodes.

3.6 Battery SOC Dynamics

SOC evolves with round-trip efficiency correction. During charging ($P_{\beta at} \geq 0$):

$$\Delta\mathrm{SOC} = \frac{P_{bat,t} \cdot \eta_C \cdot \Delta t}{E_{bat}} \quad (5)$$

During discharging ($P_{\beta at} < 0$):

$$\Delta\mathrm{SOC} = \frac{P_{bat,t} \cdot \Delta t}{\eta_D \cdot E_{bat}} \quad (6)$$

where ηc = 0.95, ηd = 0.90, Δt = 1h, $E_{βat}$ = 20.0 kWh. Updated SOC is clipped to [0.10, 0.90].

Open-Circuit Voltage Model

Terminal voltage links directly to SOC:

$V_t = V_{min} + SOC_t \cdot (V_{max} - V_{min})$

with Vmin = 49.3 V, Vmax = 62.01 V. This lets the agent predict voltage consequences of its dispatch decisions one step ahead — preventing overvoltage rather than reacting to it.

Renewable Energy Ratio

Episode-level RER over T = 24 hourly steps:

$$\mathrm{RER}_{ep} = \min\left(\frac{\sum_{t=0}^{T-1} e_{served,t}}{\max(\sum_{t=0}^{T-1} P_{load,t} \cdot \Delta t,\ \varepsilon)} \times 100\%,\ 100\%\right) \quad (7)$$

where

$$e_{served,t} = \max\left(0,\ P_{load,t} - P_{import,t}\right) \cdot \Delta t \quad (8)$$

SAC Update Equations

Bellman backup target using clipped double-Q:

$y_t = r_t + \gamma(1 - d_t)[\min(Q_1'(s_{t+1}, \tilde{a}),\ Q_2'(s_{t+1}, \tilde{a})) - \alpha \log \pi_\theta(\tilde{a}|s_{t+1})]$ (9)

Actor maximises entropy-augmented Q-value:

$\mathcal{L}_A(\theta) = \mathbb{E}_{s_t \sim \mathcal{D},\ \tilde{a} \sim \pi_\theta}[\alpha \log \pi_\theta(\tilde{a}|s_t) - \min(Q_{\phi_1}(s_t, \tilde{a}),\ Q_{\phi_2}(s_t, \tilde{a}))]$ (10)

with γ = 0.99, α = 0.20, soft target update τ = 0.005

IV. Results and Discussions

4.1 Dataset Characteristics

In Figure 1, the entire dataset is depicted over its 7,710-record length in terms of hourly PV generation (top), aggregate load (middle), and SOC distribution (bottom). PV generation is very right skewed — mean 0.513 kW, peak 8.279 kW, and median only 0.029 kW since the hours of night time and overcast are much more frequent. Generation is focused between 07:00 and 17:00, with the peak mean output of 2.731 kW at hour 8. Residential load is relatively even: mean 0.240 kW, peak 1.058 kW The 2.1:1 mean PV-to-load ratio proves that the system is generation surplus overall so morning ramp control of oversupply only exceeds reactive power profitability and tends to be a major control issue instead of supply deficit.

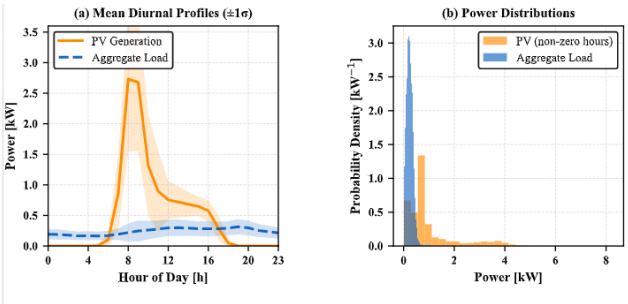

Fig. 1: Dataset characteristics. (a) Mean hourly PV generation — peak 2.731 kW at hour 8, zero overnight. (b) Mean hourly aggregate load — flat profile, mean 0.240 kW. (c) Battery SOC distribution across 7,710 records spanning full 0.10–0.90 range.

4.2 Training Convergence

Episode-by-episode reward and RER of SAC and PPO (20 training episodes) are shown in Figure 2. Rolling up from episode 1 to the end, SAC did three violations — early in episode 1. Why the episode-1 violations occur: because you are not employing a warm-start agent or one with high exploration rate → note that in the beginning the replay buffer is almost empty and so it behaves similar to random. Until enough experience has been created to give meaningful updates. Including the voltage constraint drops from episode 2, policy consistently satisfies the voltage constraint with train mean RER of 93.32 (std 11.86pp).

PPO is a different story. In episode 1, it logs 2 violations and log12 violations in the pitiful ep4 with only 26.67% RER. During training, mean RER drops alarmingly low for SAC to 93.32% compared to a mean below 70.52%, while PPO never gets above the mark during training with an all time high of episode 20 going at 88.47%. This last one is the fundamental reason for this not working: where PPO updates using all experience following an on-policy update, the rare important transitions involving overvoltage that encode the most helpful learning for avoiding voltage constraint violations occur once and are discarded after each update. The rule-based controller (flat reference line) stays at 91.87% mean RER (std 2.55pp). A trustworthy but inherently limited behaviour.

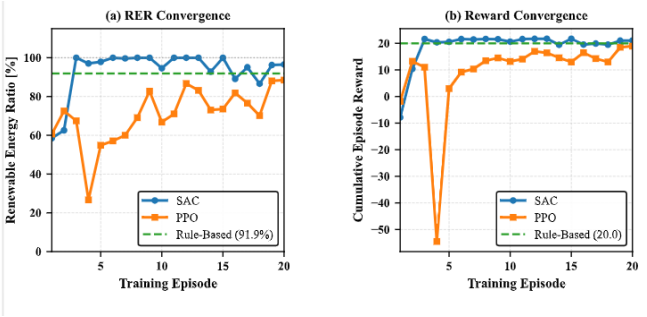

Fig. 2 caption: Training convergence over 20 episodes. SAC eliminates violations by episode 2 and reaches 100% RER at episode 3. PPO records 10 violations in episode 4 and never exceeds 88.47% RER during training. Rule-based flat reference at 91.87%.

4.3Test Performance

The mean metrics over the entire set of five held-out test episodes are shown in Table I. Both SAC and PPO achieve zero grid import and have zero violations while reaching 100% RER — even the less energy-efficient agent (PPO) confirms that they learned to remove the renewable-voltage trade-off, which will restrict the droop controller for this scenario to 92.26 % RER and an average of 0.499 kWh Import per episode once trained.

The more interesting comparison is in the way that agents make use of the voltage operating range. The mean bus droop controller operates around a nominal bus voltage of 55.42 V — this corresponds roughly to halfway between the maximum and minimum limits for the 50–60 V window. The value of SAC = 50.80 V is therefore obtained at the exact lower limit These were NOT hard-coded — these were learned from training. But overall the agent always ensures that SOC stays low to get as close to 60 V limits at the moment that morning PV ramping arrives, so the whole surplus can be absorbed without saturating. The standard deviation of voltage for SAC (0.77 V), the lowest of all three

controllers, versus droop (1.28 V) and PPO (0.97 V) means our SAC policy is both more aggressive and also much more stable in its operation.

The larger mean voltage deviation for SAC (3.20 V from nominal) versus droop (1.49 V) is not a problem — it reflects the deliberate low-SOC strategy, and both controllers stay within the safe 50–60 V band throughout.

Table I. Test Performance — 5 Held-Out Episodes (mean ± std)

Metric	Rule-Based	SAC (Proposed)	PPO
Mean RER (%)	92.26 ± 0.68	100.00 ± 0.00	100.00 ± 0.00
Grid Import (kWh)	0.499 ± 0.066	0.000 ± 0.000	0.000 ± 0.000
Voltage Violations	0	0	0
Mean V_"bus" (V)	55.42 ± 0.12	50.80 ± 0.003	50.93 ± 0.006
Mean V Deviation (V)	1.49 ± 0.07	3.20 ± 0.002	3.12 ± 0.005
Voltage Std (V)	1.28 ± 0.12	0.77 ± 0.006	0.97 ± 0.009
Mean Reward	21.12 ± 0.12	21.59 ± 0.003	21.68 ± 0.005

4.4SOC and Voltage Trajectory

The 24-hour SOC and bus voltage profile of SAC on a representative example test episode is shown in Fig. The agent begins from SOC = 0.374 (V_bus = 54.05 V) and discharges steadily throughout the night such that by hour 3, it reaches the minimum bound of SOC=0.10 (V_bus=50.57 V). The entire rest of the episode it is in a flat SOC profile that summarizes the information in the learned policy. The agent allows the minimum state of charge early in order to leave most of the charging headroom towards morning for PV ramp hours 8–15. You will never get the battery even close to SOC = 0.90 ceiling (V = 62.01 V) that will trip over voltage Each of the 24 steps has zero violations, and no grid import – all load served locally throughout.

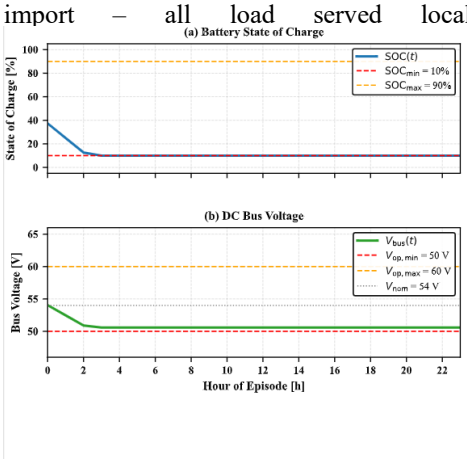

Fig. 3 caption: SAC test episode SOC and voltage trajectory. Agent drains to SOCmin = 0.10 by hour 3 and holds there — maximising charging headroom for morning PV ramp. Zero violations, zero grid import across all 24 steps.

4.5 Computational Performance

Note that SAC training takes about 82 seconds to complete on regular CPU hardware w/o GPU accelerated. PPO takes about 71 seconds. Evaluation on the basis of rules runs in less than 2 seconds. The combined test evaluation across all three controllers is done in 10 seconds max. In this section, we will discuss some of the resources required for SAC replay buffer in a 6 dimensional state space (< 5MB) A single forward pass through the two-layer 64-unit actor network provides decision latency in sub-milliseconds at inference time, fast enough for real-time hourly control on embedded hardware.

Conclusion

In this work, we proposed a SAC-based autonomous voltage control framework for isolated PV-battery microgrids validated against one full year of real operational data from the Mafate microgrid in Réunion Island. The most important design decision, namely tying SOC to bus voltage via a linear open-circuit state model, is what provides the agent with voltage-predictive state information allowing it to proactively manage SOC rather than reactively responding afterwards when violations have occurred.

The results are clear. SAC obtained 100% RER and zero grid import over all five held-out test episodes, while the rule-based droop controller achieved 92.26% RER and grid import of 0.499 kWh/episode. However, the most interesting part of the learned policy was not programmed — it emerged directly from training. With an early draining to SOCmin = 0.10 in each episode, SAC maximizes charging headroom for the morning PV ramp by fully absorbing the surplus while remaining far from the celSi voltage overvoltage ceiling of 60 V. This results in a mean bus voltage of 50.80 V and a voltage standard deviation of 0.77 V; the lowest among the three controllers tested.

Comparison with PPO indicates that SAC is the most appropriate algorithm for this task. On-policy methods squander those training transitions in a sparse-violation environment: PPO has logged 10 violations at episode 4, and the highest RER during training never surpassed 88.47%. SAC can reuse rare transitions that focus on high-value areas from a replay buffer off-policy, achieving violation-free 100% RER until episode 3.

Ongoing research will investigate automatic entropy temperature tuning to generalise across episodes sampling from differing generation profiles, application of the single-battery formulation for multi-agent coordination on distribution feeders with longer lengths and test transfer learning between microgrids operating under different PV capacities and battery technologies as a means to alleviate per-site data requirements required during deployment.

REFERENCES

[1] M. M. Kamal, “Optimal planning of standalone rural microgrid with effective dispatch strategies and battery storage,” Energy Storage, vol. 6, p. e70092, 2024, doi: 10.1002/est2.70092.

[2] N. M. Nasab and S. Yazdanian, “The advantages of lead-acid battery for off-grid design,” Energy Storage, vol. 6, p. e595, 2024, doi: 10.1002/est2.595.

[3] M. Laimon and T. Yusaf, “Towards energy freedom: Exploring sustainable solutions for energy independence in remote areas,”

Renew. Energy, vol. 222, p. 119948, 2024, doi: 10.1016/j.renene.2024.119948.

[4] M. Saleem, S. Saha, U. Izhar, and L. Ang, "Optimized energy management of a solar battery microgrid: An economic and technical perspective," J. Energy Storage, vol. 90, p. 111876, 2024, doi: 10.1016/j.est.2024.111876.

[5] A. Islam, A. Jawad, and N. Masood, "Optimal dispatch strategy of battery energy storage system in utility distribution network," J. Energy Storage, vol. 95, p. 112562, 2024, doi: 10.1016/j.est.2024.112562.

[6] C. Schade and R. Egging-Bratseth, "Battery degradation: Impact on economic dispatch," Energy Storage, vol. 6, p. e588, 2024, doi: 10.1002/est2.588.

[7] W. Wang, X. Zhang, Z. Chen, X. Zhang, Z. Li, and L. Li, "AC frequency/battery power consensus control for stand-alone modular microgrid," J. Energy Storage, vol. 118, p. 116204, 2025, doi: 10.1016/j.est.2025.116204.

[8] I. Bekki, H. Rizki, F. E. Lamzouri, E. Boufounas, and L. Bejjit, "Robust power control and energy management of a stand-alone PV-wind-battery microgrid," J. Energy Storage, vol. 169, p. 122522, 2026, doi: 10.1016/j.est.2026.122522.

[9] A. Ajagekar, B. Decardi-Nelson, and F. You, "Energy management for demand response in networked greenhouses with deep reinforcement learning," Appl. Energy, vol. 355, p. 122349, 2024, doi: 10.1016/j.apenergy.2023.122349.

[10] G. Li and S. W. Or, "Multi-agent deep reinforcement learning-based multi-time scale energy management for microgrid with hydrogen energy storage," J. Clean. Prod., vol. 466, p. 142842, 2024, doi: 10.1016/j.jclepro.2024.142842.

[11] Y. Li, J. Hou, and G. Yan, "Exploration-enhanced multi-agent reinforcement learning for distributed energy management," Appl. Energy, vol. 359, p. 122744, 2024, doi: 10.1016/j.apenergy.2024.122744.

[12] L. Ding, Y. Cui, G. Yan, Y. Huang, and Z. Fan, "Distributed energy management of multi-area integrated energy systems using multi-agent deep reinforcement learning," Int. J. Electr. Power Energy Syst., vol. 157, p. 109867, 2024, doi: 10.1016/j.ijepes.2024.109867.

[13] S. Seyednouri, A. Safari, A. Quteishat, M. Younis, and M. Saleh, "Stochastic energy management of a multi-microgrid system with battery storage," Renew. Energy Focus, vol. 48, p. 100531, 2024, doi: 10.1016/j.ref.2023.100531.

[14] F. Tao, M. Wu, and Y. Cao, "Generalized maximum entropy reinforcement learning via reward shaping," IEEE Trans. Artif. Intell., vol. 5, pp. 1563–1572, 2024, doi: 10.1109/tai.2023.3297988.

[15] X. Liu and Z. Gou, "Occupant-centric HVAC and window control: A reinforcement learning model for energy efficiency and thermal comfort," Build. Environ., vol. 250, p. 111197, 2024, doi: 10.1016/j.buildenv.2024.111197.

[16] H. Zhang and X. Han, "Off-policy asymptotic and adaptive maximum entropy deep reinforcement learning," Int. J. Mach. Learn. Cybern., vol. 16, pp. 2417–2429, 2024, doi: 10.1007/s13042-024-02399-7.

[17] D. Calogine, J. Francou, C. Abbezzot, and T. Fanjirindratovo, "Data in experimental stand-alone microgrid: Solar production, domestic loads, battery storage and meteorological series," Data Brief, vol. 51, p. 109643, 2023, doi: 10.1016/j.dib.2023.109643.

[18] H. Momen and S. Jadid, "A novel microgrid formation strategy for resilience enhancement considering battery energy storage systems," J. Energy Storage, vol. 100, p. 113565, 2024, doi: 10.1016/j.est.2024.113565.

[19] G. Maslak and P. Orłowski, "Operational optimisation of a microgrid using non-stationary hybrid switching control," Renew. Sustain. Energy Rev., vol. 202, p. 114685, 2024, doi: 10.1016/j.rser.2024.114685.

[20] K. Pang, J. Zhou, S. Tsianikas, D. Coit, and Y. Ma, "Long-term microgrid expansion planning with resilience and environmental considerations," Renew. Sustain. Energy Rev., vol. 191, p. 114068, 2024, doi: 10.1016/j.rser.2023.114068.

[21] T. T. Nguyen, T. T. Nguyen, and H. P. Nguyen, "Optimal operation of battery energy storage system in microgrid to minimize cost," J. Energy Storage, vol. 131, p. 115904, 2025, doi: 10.1016/j.est.2025.115904.

[22] Y. Wang and V. Vittal, "Data driven real-time dynamic voltage control using decentralized execution," IEEE Open J. Power Energy, vol. 11, pp. 508–519, 2024, doi: 10.1109/oajpe.2024.3459002.

[23] B. Zhang and C. He, "Decentralized voltage and var control of active distribution network based on multi-agent reinforcement learning," IEEE Access, vol. 13, pp. 115090–115101, 2025, doi: 10.1109/access.2025.3584129.

[24] C. Cui and Y. Liu, "A hierarchical HVAC optimal control method for reducing energy consumption and improving thermal comfort," Energy Convers. Manag., vol. 302, p. 118118, 2024, doi: 10.1016/j.enconman.2024.118118.

[25] F. Sakil and N. Aksoy, "Reinforcement learning-based multi-objective smart energy management for electric vehicles," Energy, vol. 322, p. 135475, 2025, doi: 10.1016/j.energy.2025.135475.

[26] M. Ahmadi and H. Aly, "A comprehensive review of reinforcement learning-based voltage control in smart grids," Renew. Sustain. Energy Rev., vol. 227, p. 116526, 2026, doi: 10.1016/j.rser.2025.116526.

[27] M. H. Nejati Amiri and F. Gueniat, "Explainable deep reinforcement learning for resilient and battery-aware microgrid energy management," Energy Convers. Manag., vol. 353, p. 121215, 2026, doi: 10.1016/j.enconman.2026.121215.

[28] X. Shen, J. Tang, F. Pan, B. Qian, and Y. Zhao, "Quantum-inspired deep reinforcement learning for adaptive frequency control in microgrids," Front. Energy Res., vol. 12, p. 1366009, 2024, doi: 10.3389/fenrg.2024.1366009.

[29] T. Patel and P. Chauhan, "Battery storage assisted plug-and-play of wind generator in standalone microgrid," IEEE Trans. Consum. Electron., vol. 70, no. 1, pp. 460–470, 2024, doi: 10.1109/tce.2023.3341950.