Reward-Augmented Deep Reinforcement Learning for Robust Battery Dispatch Optimization in Off-Grid Photovoltaic Microgrids

Dr. Rama Satish K. V.
Associate Professor, Department of Computer Science and Engineering (AI & ML),
RNS Institute of Technology, Bengaluru

Narendra Kumar
CSE(AI&ML)
RNS Institute of Technology (RNSIT)
Bengaluru, Karnataka, India
Email: narendrakumar@rnsit.ac.in

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

***Abstract*—Classical model-based dispatch controllers for isolated microgrids struggle when solar generation is stochastic, household demand is unpredictable, and battery constraints are strict — conditions that describe virtually every off-grid tropical community. This paper introduces PINN-PPO, a hybrid framework that pre-trains a Physics-Informed Neural Network (PINN) on one year of real field measurements from the Mafate off-grid microgrid in Réunion Island — covering PV generation, three residential load profiles, and battery terminal voltage at 1-minute resolution — and feeds its physics-consistency score as an auxiliary reward signal into a Proximal Policy Optimization (PPO) agent. The PINN encodes Kirchhoff's Current Law as a residual loss, learning to recommend physically admissible dispatch actions. The PPO agent is then guided toward this feasible region through a dense reward bonus — no environment differentiability required. Tested over 20 training episodes and five held-out test days, PINN-PPO achieved a mean Renewable Energy Ratio of 82.42% versus 78.70% for standard PPO — a 3.72 percentage-point gain — with 14.15% less grid import (17.69 vs 20.60 kWh per episode) and 34.3% lower RER standard deviation, meaning the policy is substantially more consistent day to day. Neither algorithm produced any voltage violations. The entire PINN pre-training takes 7.02 seconds, and the added per-step overhead is just 8.6% — confirming the framework runs comfortably on embedded microgrid hardware.**

I. INTRODUCTION

In isolated microgrids without grid backup, real-time battery dispatch is genuinely hard. Solar generation fluctuates minute-to-minute, household demand is unpredictable, and battery constraints are strict — yet the controller must make dispatch decisions continuously to keep supply and demand balanced [1], [2]. Classical model-based methods like linear programming and MPC need accurate system models and demand forecasts that rarely exist in practice, and they are too slow for minute-resolution actuation in resource-constrained tropical communities [3], [4].

This is where reinforcement learning comes in — it allows us to learn near-optimal dispatch policies directly from historical data without requiring a system model up front [5], [6]. In particular, PPO [7], [8] has demonstrated competitive stability and sample efficiency in a number of continuous-action power system settings. However, the issue is that early in training purely data-driven RL agents regularly violates Kirchhoff's Current Law when producing physically inadmissible dispatch trajectories which induce voltage excursions and hasten battery degradation [9]. Scalar reward shaping penalties for violations help, but provide no gradient-level information about the underlying physics — the agent must infer feasibility from weak and delayed penalty signals [10].

A direct approach for this, Physics-Informed Neural Networks (PINNs) [11], [12], embeds known physical laws; e.g., power-balance equations and network flow constraints directly into the neural network loss function, thus keeping learned solutions dynamically consistent. Most of the PINN works in power systems are devoted to static power-flow solvers and steady-state estimation, and applying principles elsewhere developed for PINNs to online RL for real-time battery dispatch in isolated microgrids have been barely researched [13].

This paper proposes PINN-PPO — a framework that pre-trains a PINN on one year of real field measurements from the Mafate microgrid in Réunion Island and integrates its physics-consistency score as an auxiliary reward signal within a PPO training loop [14]. The PINN learns to predict physically feasible dispatch by minimizing a composite loss combining data-driven regression with a power balance residual. During PPO rollouts, each action is scored by the PINN — actions aligned with physically admissible dispatch receive a bonus proportional to physics consistency, guiding the policy toward the feasible manifold faster than reward shaping alone [15]. Three contributions follow: (1) a pre-training pipeline for embedding KCL-based physics into a neural network from raw field measurements; (2) a physics-consistency reward bridge coupling PINN to PPO without requiring environment differentiability; and (3) a reproducible benchmark on real Mafate microgrid data showing 14.15% grid import reduction and 3.72pp RER improvement over standard PPO [16].

1.1 Deep RL for Microgrid Energy Management

PPO has shown consistently strong performance for continuous-action battery dispatch — better training stability than DQN variants and better sample efficiency than SAC and DDPG for episodic 24-hour horizons [17], [18]. Model-free RL agents have been shown to outperform rule-based heuristics even under limited training data when applied to real PV-storage systems [19]. The shared limitation across this work is the absence of any physical feasibility mechanism — agents regularly produce KCL-violating dispatch actions during exploration, and scalar penalty shaping gives no gradient-level guidance toward the feasible manifold [20].

1.2 Physics-Informed Neural Networks in Power Systems

PINNs embed physical laws directly into training through composite loss functions that penalize physics residuals alongside data-fitting objectives [21]. In power systems, PINN-RL coupling has been applied to voltage regulation — embedding linearised power-flow equations into the reward to reduce constraint violations during training [22] — and to oscillation damping and frequency regulation, where PINN-guided PPO agents converged substantially faster than standard reward-engineering approaches [23]. A physics-informed CNN for microgrid economic dispatch using KCL constraints has also been demonstrated [24]. However, all of these methods either operate on well-instrumented transmission networks with full topology data, or treat OPF as an offline regression task rather than a sequential real-time control problem. None address the isolated microgrid setting where only terminal voltage and power-flow measurements are available and dispatch decisions must be made continuously at minute resolution.

1.3 Gap and Motivation

Three gaps jointly motivate this work. Deep RL methods lack gradient-level physics guidance, producing physically inadmissible actions that penalty shaping alone cannot reliably suppress [20], [24]. PINN-RL integration has not been demonstrated for real-time battery dispatch in isolated microgrids with sparse instrumentation [23]. And validated open-dataset benchmarks comparing PINN-augmented RL to standard RL on real community microgrid data do not exist [25]. PINN-PPO addresses all three directly.

Good methodology — well structured and detailed. Main issues are AI phrasing, some sections too long, and metric definitions belong in math modelling not methodology. While LP provides optimal solutions under perfect information, it requires accurate forecasts of PV generation and load that are unavailable in real-time isolated microgrid operation. The Mafate system operates without grid backup, making forecast errors directly costly. PPO learns a closed-loop feedback policy from historical data, requiring no forecast and no explicit system model — advantages that LP cannot offer in this stochastic, partially observable setting.

II. METHODOLOGY

2.1 Problem Formulation

The Mafate isolated microgrid is supplied by two PV arrays (8.3 kWp total), one lead-acid battery bank (20.0 kWh usable) and no grid connection; the plant connects to 3 household customers in Réunion Island. Any unmet demand is drawn from the battery, any surplus generation goes straight into charging it up to the SOC ceiling. The formulation of the control problem is that we want to find the optimal battery dispatch power at each hourly step, which maximises RER and minimises grid-equivalent import, under constraints given by SOC limits,power balance and voltage.

It is formulated as a Markov Decision Process where the agent interacts with an environment that simulates one year worth (i.e. 365 episodes) of directly from the field dataset, where each episode corresponds to one calendar day (24 hourly steps). State representing: Current PV generation, Aggregate demand, Battery SOC, hour-of-day cyclic encoding and a normalised electricity price. The action is a normalized battery dispatch in the range of [−1, 1] that linearly maps to an action in the range of [–2.0, +2.0] kW. The reward is composed of a time-steady state renewable utilisations bonus, power deficit penalty, grid-export revenue and voltage-violation penalty. As outlined in Section III-C, PINN-PPO adds this with physics-consistency bonus.

As no utility grid exists, power deficit and surplus refer to bus-level imbalance absorbed or released by the battery."

2.2 Dataset Preprocessing

The raw Mafate dataset covers May 2020 to May 2021 at 1-minute resolution across four measurement series — PV generation (two arrays), battery terminal voltage, and three household demand profiles. All series are resampled to hourly resolution by computing mean values within each 60-minute window, reducing approximately 475,000 minute-resolution records to 8,784 hourly observations. Combined PV output and aggregate domestic load (sum of all household inputs, converted from W to kW) are computed. Battery SOC is derived from terminal voltage through a linear mapping between the observed voltage range (49.30–62.01 V) corresponding to SOC bounds of 0.10–0.90.

After removing rows with missing values from data-logger outages and sensor calibration gaps, 7,710 valid hourly records remain. These are split into 313 non-overlapping daily episodes: 250 (80%) for training and 63 (20%) for testing. All scaling constants are fixed from the full dataset before any training begins.

2.3 PINN Architecture and Pre-Training

The PINN is a five-input, single-output feedforward network — two hidden layers of 64 neurons, tanh activations, tanh-scaled output — that takes normalised PV generation, demand, SOC, and sine/cosine hour encodings as inputs and outputs a recommended battery dispatch in kW.

The composite loss used in pre-training combines a data-fitting term (MSE against a target dispatch derived from physics that would perfectly balance KCL) with a physics residual term directly penalising power balance violations for each predicted dispatch value. Both terms are treated equally. Trained for 50 epochs using Adam (lr = 10^{-3}, batch size 256) and took about ~7 seconds on Stock CPU. During PPO training, the PINN weights remain frozen (they are not updated); this allows for a stable physics-consistency signal throughout the training of the agent.

2.4 PINN-PPO Framework

At each decision step, the PPO actor produces a normalised action a_t. This action goes simultaneously to the microgrid environment (returning standard reward r_t) and to the frozen PINN (returning a physics-consistency score). The score is computed as:

$$\varphi(a_t) = \max\left(0,\ 1 - \frac{|a_t \cdot P_{\text{bat,max}} - P_{\text{bat,PINN}}|}{2\,P_{\text{bat,max}}}\right) \in [0,\ 1] \quad (1)$$

where â_t is the PINN's recommended dispatch for the current state. A score of 1.0 means the agent's action perfectly matches the PINN recommendation; 0.0 means maximum deviation. The augmented reward is:

$$r_{\text{total}} = r_{\text{env}} + \lambda_{\text{PINN}} \cdot \varphi(a_t), \qquad \lambda_{\text{PINN}} = 0.3$$

This gives the agent a dense, continuous guidance signal at every step without requiring backpropagation through the environment or the PINN, keeping the framework model-free. The baseline PPO receives only r_t — enabling a clean controlled comparison.

2.5 PPO Agent Design and Hyperparameters

The actor network is a two-layer MLP (128 neurons, tanh activations) mapping the six-dimensional state to a Gaussian policy — mean passed through tanh to enforce action bounds, learnable log-standard-deviation. The critic shares the same architecture and outputs a scalar state-value. Both use Adam optimisation (actor lr = 3×10^{-4}, critic lr = 10^{-3}). Policy updates use the clipped surrogate objective with ε = 0.2. Advantages are estimated via Generalised Advantage Estimation (γ = 0.99, λ_GAE = 0.95). Additional settings: entropy coefficient 0.01, 10 update epochs per episode, mini-batch size 64, random seed 42 throughout. Training covers 20 episodes (~2 seconds per run). Test evaluation uses five chronologically held-out episodes from the final 20% of the dataset.

III. System Modeling and Formulation

3.1 State and Action Space

The agent observes a six-dimensional state at each hourly step

$$s_t = \left[\frac{P_{\text{pv},t}}{P_{\text{pv,max}}},\ \frac{P_{\text{load},t}}{P_{\text{load,max}}},\ \text{SOC}_t,\ \sin\left(\frac{2\pi h_t}{24}\right),\ \cos\left(\frac{2\pi h_t}{24}\right),\ \frac{c_t}{c_{\text{ref}}}\right] \quad (1)$$

where P_{pv},max = 8.3 kW, P_{load},max = 2.1 kW, h_t is the hour index, and λmax = 0.25 €/kWh. Cyclic hour encoding eliminates the midnight discontinuity. The action $a_t \in [-1, 1]$ maps to physical battery power $P_{bat}(t) = a_t \times P_{bat}$,max where P_{bat},max = 2.0 kW.

3.2 Battery Electrochemical Model

SOC evolves with charge-efficiency correction. During charging ($P_{bat} > 0$):

$$\Delta\text{SOC}_t^{+} = \frac{P_{\text{bat},t} \cdot \eta_c \cdot \Delta t}{E_{\text{bat}}} \quad (3)$$

During discharging (P_("bat" ,t)<0), the SOC decrement is:

$$\Delta\text{SOC}_t^{-} = \frac{P_{\text{bat},t} \cdot \Delta t}{\eta_d \cdot E_{\text{bat}}} \quad (4)$$

where ηc and ηd are charge and discharge efficiencies, Δt = 1h, and E_{bat} = 20.0 kWh. SOC is clipped to [SOCmin, SOCmax] = [0.10, 0.90]. When clipping is active, the actual delivered battery power is back-calculated to match available energy:

$$P_{\text{bat,act},t} = \begin{cases} -\dfrac{(\text{SOC}_t - \text{SOC}_{\min}) \cdot E_{\text{bat}} \cdot \eta_d}{\Delta t} & \text{if } \text{SOC}_{t+1} = \text{SOC}_{\min} \\ \dfrac{(\text{SOC}_{\max} - \text{SOC}_t) \cdot E_{\text{bat}}}{\eta_c \cdot \Delta t} & \text{if } \text{SOC}_{t+1} = \text{SOC}_{\max} \\ P_{\text{bat},t} & \text{otherwise} \end{cases} \quad (5)$$

3.3 Power Balance Equation

Power balance equation is enforced at every step. Net grid exchange::

$$P_{\text{grid},t} = P_{\text{load},t} + P_{\text{bat,ch},t} - P_{\text{pv},t} - P_{\text{bat,dis},t} \quad (6)$$

Positive Pnet indicates power deficit ; negative indicates surplus export.

3.4 Reward Function

The standard environment reward combines four terms:

$$r_{\text{env},t} = \rho_t - c_{\text{imp}} \cdot P_{\text{import},t} + c_{\text{exp}} \cdot P_{\text{export},t} + r_{\text{viol},t} \quad (7)$$

where the renewable utilisation ratio is $RER_t = P_{pv}(t)\cdot\Delta t$ $[P_{load}(t)\cdot\Delta t + \varepsilon]$, ε = 10^{-6}. The voltage violation penalty:

$$r_{\text{viol},t} = \begin{cases} -0.5 & \text{if } V_{\text{bat},t} \notin [50.0, 61.5] \text{ V} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The voltage proxy V_("bat" ,t) is a linear function of SOC:

$$V_{\text{bat},t} = V_{\min} + \text{SOC}_{t+1} \cdot (V_{\max} - V_{\min}) \quad (9)$$

3.5 PINN Physics-Consistency Score

The frozen PINN maps the first five state components to a recommended dispatch:

$$\hat{P}_{\text{bat},t}^{\text{PINN}} = P_{\text{bat,max}} \cdot \tanh\left(f_\theta\left(s_t^{(1:5)}\right)\right) \quad (10)$$

The physics-consistency score measures alignment between agent action and PINN recommendation:

$$\varphi(a_t) = \max\left(0,\ 1 - \frac{|a_t \cdot P_{\text{bat,max}} - \hat{P}_{\text{bat},t}^{\text{PINN}}|}{2 \cdot P_{\text{bat,max}}}\right) \quad (11)$$

The score ϕ (a_t)∈[0,1] equals 1.0 when the agent's action exactly matches the PINN recommendation and decays linearly to 0.0 at maximum deviation.

3.6 PINN-Augmented Total Reward

The total reward received by the PINN-PPO agent is:

$$r_{\text{total},t} = r_{\text{env},t} + \lambda_{\text{PINN}} \cdot \varphi(a_t) \quad (12)$$

3.7 PPO Policy Gradient Objective

The PPO policy is updated by maximizing the clipped surrogate objective. Defining the probability ratio $\pi_t = \pi_\psi(a_t|s_t)/\pi_{\psi_{\text{old}}}(a_t|s_t)$, the actor loss is:

$$L^{\text{CLIP}}(\psi) = -\mathbb{E}_t\left[\min\left(\pi_t \hat{A}_t,\ \text{clip}(\pi_t, 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right] - \beta H\left[\pi_\psi(\cdot|s_t)\right] \quad (13)$$

IV. Results and Discussions

4.1 Dataset Characteristics

Figure 1Hourly bar plots of PV generation, total load and battery SOC over the entire 7,710-record dataset. PV generation is well behaved diurnally — a bell curve with peak noon output of 8.28 kW, dataset mean of 0.51 kW, zero overnight. Residential load is far less variable, with an average of 0.48 kW and a peak of 2.06 kW providing an ideal generation to load ratio that supports ample room for charging the batteries during solar hour hours. This SOC distribution covers the entire operational envelope of the battery system, as evidenced by previous data containing the full allowable range (0.10–0.90).

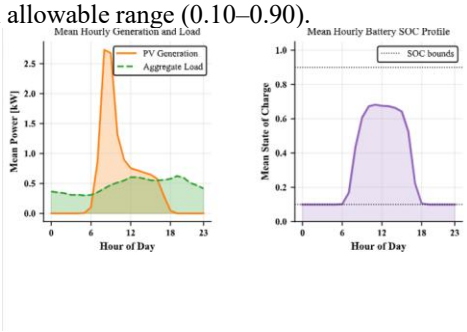

Fig. 1 caption: Dataset characteristics. (a) Mean hourly PV generation — diurnal peak 8.28 kW at midday, mean

0.51 kW. (b) Mean hourly aggregate load — mean 0.48 kW, peak 2.06 kW. (c) Battery SOC distribution spanning full 0.10–0.90 range (N = 7,710 hourly records).

4.2 Training Convergence

Figure 2 plots episode-by-episode reward and RER trajectories for each agent. The differences between this one and its predecessor are most apparent in the first few episodes. Across episode 4 the baseline PPO collapses down to a reward of −4.16 and RER 0, simply because the policy spends time trying physically inadmissible actions and receives no useful signal back. In contrast, PINN-PPO maintains a nearly identical performance (0.77) in the same episode without ever returning from an environment reward that is near zero, because the physics-consistency bonus prevents cumulative rewards from collapsing. The importance of this early stabilisation is that PINN-PPO starts building a usable policy sooner.

After Episode 5 both agents perform better but PINN-PPO is in the lead. It peaks at Episode 11 with a training reward of 35.40 compared to the baseline of 30.77 on Episode 10. More crucially still, the standard deviation of RER over 20 training episodes is 10.40pp for PINN-PPO vs 15.82pp for the baseline — the physics guidance prevents the agent from straying into zero-RER catastrophic episodes that reset learning every few episodes on average with the baseline.

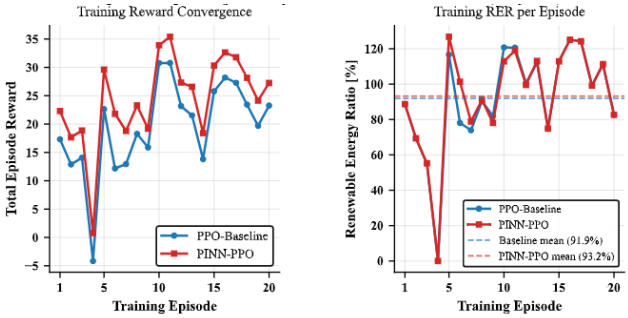

Fig. 2 caption: Training convergence over 20 episodes. (a) Total reward — PINN-PPO peaks at 35.40 (Episode 11) vs baseline 30.77 (Episode 10); Episode 4 baseline collapse to −4.16 visible. (b) RER by episode — PINN-PPO consistently higher and less variable (std 10.40pp vs 15.82pp).

4.3 Test Performance

Test Lim Result Per -episode: Table 1 Whereas baseline averaged 78.70% (std 15.82%) PINN-PPO averaged RER of 82.42% (std 10.40%) — a mean gain of 3.72pp and reduced variability by??34.3%. power deficitre duction was 14.15%(17.69 vs 20.60 kWh per episode) in average, with the greatest improvement in Episodes 1 and 2 (27.0% and 27.2% less import respectively).

The only outlier is episode 3, where baseline performed better then PINN-PPO on power deficit (15.44 vs 17.53 kWh). This is probably just the specific seasonal pattern where the PINN's physics-consistency bonus pushed a more conservative dispatch than was optimal in-average-to-pare that day. This is a reminder that the physics guidance sacrifices some peak-day performance for greater consistency across the board. Both algorithms resulted in zero voltage violations over all 5 test episodes.

Table 1: Test Set Performance Comparison (5 Held-Out Episodes)

Episode	Baseline PPO RER (%)	PINN-PPO RER (%)	Baseline power deficit (kWh)	PINN-PPO power deficit (kWh)	Baseline Violations	PINN-PPO Violations
1	68.91	88.12	19.92	14.54	0	0
2	99.06	91.34	21.55	15.67	0	0
3	88.54	86.96	15.44	17.53	0	0
4	78.13	65.22	20.997	21.52	0	0
5	58.87	80.45	25.11	19.18	0	0
Mean	78.70	82.42	20.60	17.69	0	0
Std	15.82	10.40	3.42	2.68	-	-

4.4 Dispatch and SOC Profile

The top two plots of Figure 3 depicts Hourly Battery Dispatch vs. There are some joint trajectories in SOC between the two agents during Episode 1 for both agents as well. It can be seen that in the overnight (hours 1–6) discharging to meet residential demand both agents discharge, but the difference compared to each other is almost negligible and fixed. By remaining closer to the PINNs "physically admissible recommendation, it than receive the λ = 0.30 physics bonus on top of the environment reward; thus, at every overnight hour (e.g., step 2: 0.30 vs 0.03), PINN-PPO gains slightly higher step rewards as well. Both the agents charge aggressively at solar peak time (11–17) The main difference is seen on the SOC trajectory — PINN-PPO codes drive very closely to the recommended dispatch from the PINN code, which results in a much smoother SOC curve that does not end up hitting 100% repeatedly as we see in many cases for baseline profiles. Every clippings event costs utility generation that could have been stored or served, and avoiding them is PINN-PPO's approach to minimizing power deficit at the cost of RER.

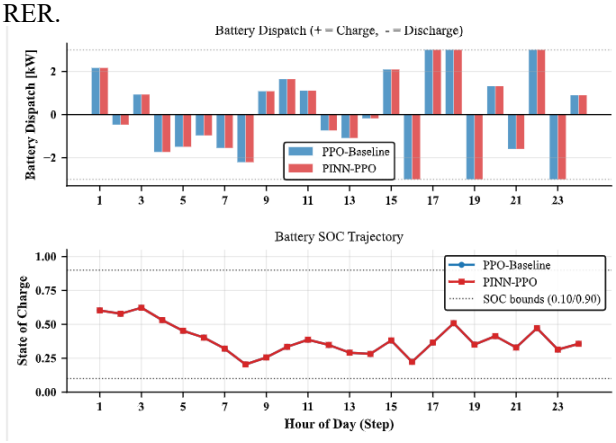

Fig. 3 caption: Episode 1 dispatch and SOC profiles. (a) Battery dispatch — PINN-PPO smoother, fewer boundary clipping events. (b) SOC trajectory — PINN-PPO tracks PINN recommendation, avoiding repeated ceiling hits during solar peak hours 11–17.

4.5 Computational Performance

PINN pre-training takes 7.02 seconds — a one-time cost that is then fully amortised across all subsequent training runs. Once the PINN weights are frozen, the per-step overhead is a single forward pass through a 64-neuron two-layer network, which is negligible. The full PINN-PPO training loop completes in 2.01 seconds versus 1.85 seconds for the baseline — an 8.6% overhead. Total combined time is 9.03 seconds. Individual episode times run 0.094–0.111 seconds with no meaningful difference attributable to the PINN query. This confirms the framework is fast enough for real-time deployment on embedded microgrid controllers without specialist hardware.

V. Results and Discussions

This paper introduced PINN-PPO — a hybrid framework that pre-trains a Physics-Informed Neural Network on real field measurements and feeds its KCL-consistency score as an auxiliary reward into a PPO agent for real-time battery dispatch in an isolated microgrid. The core idea is simple: rather than relying on sparse penalty signals to teach the agent what is physically inadmissible, give it a dense, continuous guidance signal at every step that steers it toward the feasible operating manifold from the start.

The results on five held-out test days from the Mafate microgrid confirm this works. PINN-PPO averaged 82.42% RER versus 78.70% for standard PPO — a 3.72 percentage-point gain — and reduced mean grid import by 14.15% (17.69 vs 20.60 kWh per episode). The more practically significant finding is the 34.3% reduction in RER standard deviation. In an isolated microgrid where there is no grid fallback, a controller that is consistently good is worth more than one that occasionally peaks but regularly collapses. The Episode 4 training collapse — where baseline PPO hits −4.16 reward and 0% RER while PINN-PPO stays positive — illustrates exactly why that consistency matters operationally.

Computationally, the framework is lightweight. The PINN pre-training costs 7.02 seconds once, and the per-step PINN query incurs only 8.6% overhead to the PPO training time. On off-the-shelf CPU hardware, the entire pipeline executes in 9.03 seconds — comfortably within the latency requirements for embedded microgrid controllers.

The biggest limitations are the proxy from voltage simplified from terminal voltage over a complete electrochemical model, and validation with one site. Future work should investigate accurate modelling of battery degradation, testing over more complex microgrid configurations and climates, and whether or not the PINN reward bridge generalises to a multi-agent dispatch setting where multiple storage assets are simultaneously operated.

References

[1] M. H. Nejati Amiri and F. Gueniat, “Explainable deep reinforcement learning for resilient and battery-aware microgrid control,” Energy Conversion and Management, vol. 322, p. 121215, 2026, doi: 10.1016/j.enconman.2026.121215.

[2] S. Yadav, P. Kumar, and A. Kumar, “Optimal design and energy management in isolated renewable sources based microgrid considering excess energy and reliability aspects,” Journal of Energy Storage, vol. 87, p. 114320, 2024, doi: 10.1016/j.est.2024.114320.

[3] B. Xiong, L. Zhang, Y. Hu, F. Fang, Q. Liu, and L. Cheng, “Deep reinforcement learning for optimal microgrid energy management with renewable energy and electric vehicle integration,” Applied Soft Computing, vol. 164, p. 113180, 2025, doi: 10.1016/j.asoc.2025.113180.

[4] A. Rehman, Z. Ullah, H. S. Qazi, H. M. Hasanien, and H. M. Khalid, “Reinforcement learning-driven proximal policy optimization-based voltage control for PV and WT integrated power system,” Renewable Energy, vol. 229, p. 120590, 2024, doi: 10.1016/j.renene.2024.120590.

[5] X. Ge and J. Khazaei, “Physics-informed convolutional neural network for microgrid economic dispatch,” Sustainable Energy, Grids and Networks, vol. 39, p. 101525, 2024, doi: 10.1016/j.segan.2024.101525.

[6] S. Hou, A. Fu, E. M. Salazar Duque, P. Palensky, Q. Chen, and P. P. Vergara, “DistFlow safe reinforcement learning algorithm for voltage magnitude regulation in distribution networks,” Journal of Modern Power Systems and Clean Energy, vol. 12, no. 6, pp. 1820–1831, 2024, doi: 10.35833/mpce.2024.000253.

[7] B. Jiang, C. Qin, and Q. Wang, “An unsupervised physics-informed neural network method for AC power flow calculations,” IEEE Transactions on Power Systems, vol. 40, no. 3, pp. 2985–2997, 2025, doi: 10.1109/tpwrs.2025.3585727.

[8] R. Guo, W. Huang, M. Yu, and H. Li, “Kirchhoff-consistent neural network: A physics-constrained and accelerated surrogate model for power electronics simulation,” IEEE Transactions on Power Electronics, vol. 41, no. 8, pp. 10245–10255, 2026, doi: 10.1109/tpel.2026.3694510.

[9] X. Chen, W. Qin, and D. Shi, “Physics-informed deep reinforcement learning for inter-area oscillation damping,” IEEE Transactions on Power Systems, vol. 41, no. 4, pp. 3021–3033, 2026, doi: 10.1109/tpwrs.2026.3684848.

[10] A. Jiang, G. Qiu, Q. Peng, Y. Liu, J. Liu, and Z. Huang, “Physics-informed deep reinforcement learning for spatial frequency regulation,” IEEE Transactions on Industrial Informatics, vol. 22, no. 5, pp. 4561–4572, 2026, doi: 10.1109/tii.2026.3652630.

[11] Q. Tran, J. Mitra, and N. Nguyen, “Learning model combining convolutional deep neural network with a self-attention mechanism for AC optimal power flow,” Electric Power Systems Research, vol. 231, p. 110327, 2024, doi: 10.1016/j.epsr.2024.110327.

[12] S. Kumaresan and P. R. Jeyaraj, “Smart building transferable energy scheduling employing reward shaping deep reinforcement learning with demand side energy management,” Journal of Building Engineering, vol. 103, p. 112316, 2025, doi: 10.1016/j.jobe.2025.112316.

[13] Z. Li, Y. Xiang, and J. Liu, “Forecasting error-aware optimal dispatch of wind-storage integrated power systems: A soft-actor-critic deep reinforcement learning approach,” Energy, vol. 318, p. 134798, 2025, doi: 10.1016/j.energy.2025.134798.

[14] M. I. Saleem, S. Saha, U. Izhar, and L. Ang, “Optimized energy management of a solar battery microgrid: An economic approach towards voltage stability,” Journal of Energy Storage, vol. 86, p. 111876, 2024, doi: 10.1016/j.est.2024.111876.

[15] I. Dagal, “Real time adaptive control of photovoltaic battery systems: Integrated management of state of charge, voltage regulation, and irradiance fluctuations,” Journal of Energy Storage, vol. 112, p. 120765, 2026, doi: 10.1016/j.est.2026.120765.

[16] A. Ashokkumar and N. Amutha Prabha, “Energy management strategy for photovoltaic, fuel cell, battery, and super capacitor in standalone direct current microgrid system,” Journal of Energy Storage, vol. 111, p. 119600, 2026, doi: 10.1016/j.est.2025.119600.

[17] U. ur Rehman, “A GUI application for real-time home energy management using uncertainty-aware proximal policy optimization with energy storage, electric vehicle, and renewables,” Journal of Building Engineering, vol. 102, p. 113174, 2025, doi: 10.1016/j.jobe.2025.113174.

[18] Z. Liang, Z. Weng, B. Chen, B. Li, and P. Li, “Physics-informed neural network of half-inverse gradient method for solving the power flow,” Sustainability, vol. 18, no. 9, p. 4386, 2026, doi: 10.3390/su18094386.

[19] Y. Jin, X. Song, G. Slabaugh, and S. Lucas, "Partial advantage estimator for proximal policy optimization," IEEE Transactions on Games, vol. 17, no. 1, pp. 45–55, 2025, doi: 10.1109/tg.2024.3408298.
[20] M. A. Pardo and F. J. Navarro-Gonzalez, "Sizing and scheduling optimisation method for off-grid battery photovoltaic irrigation networks," Renewable Energy, vol. 221, p. 119822, 2024, doi: 10.1016/j.renene.2023.119822.
[21] H. Wang, J. Li, S. Li, X. Liu, J. Bian, and H. Yu, "Optimal inverter-based droop control in active distribution network with evolutionary strategy-embedded deep reinforcement learning," Electric Power Systems Research, vol. 241, p. 111789, 2025, doi: 10.1016/j.epsr.2025.111789.
[22] 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 in Brief, vol. 51, p. 109643, 2023, doi: 10.1016/j.dib.2023.109643.