

Reinforcement Learning For Optimal Power Flow Control In Renewable-Integrated Microgrids

Md Mazharul Haque¹, Nitish Kumar², Mr. Maruti Nandan Mishra³, Prabhat Kumar⁴,
Mayank Kumar Singh⁵, Raju Rajak⁶

^{1,2,4,5}Assistant Professor, Department of Electrical and Electronics Engineering, Darbhanga College of Engineering, Darbhanga, Bihar (India)

³Assistant Professor, Department of Electrical and Electronics Engineering, Sersha Engineering College, Sasaram, Bihar (India)

⁶Assistant Professor, Department of Electrical and Electronics Engineering, Katihar Engineering College, Katihar, Bihar (India)

ABSTRACT

The rapid integration of stochastic renewable energy sources (RES) into distribution-level microgrids fundamentally undermines traditional optimisation-based optimal power flow (OPF) methodologies, which depend on precise system models and deterministic forecasts that are seldom accessible in practice. This research presents a deep reinforcement learning (DRL) framework using the Proximal Policy Optimisation (PPO) method to address the OPF problem in hybrid microgrids containing photovoltaic (PV) arrays, wind turbines, battery energy storage systems (BESS), and diesel backup generators. The proposed model learns a continuous, model-free control policy that simultaneously minimises operating costs, curtails voltage variances, and maximises renewable utilisation under time-varying load and generation uncertainty. The environment is constructed as a Markov Decision Process (MDP) utilising an IEEE 33-bus distribution test feeder enhanced with renewable generation profiles obtained from real-world irradiance and wind speed data. Comparative assessment with Model Predictive Control (MPC), Deep Q-Network (DQN), and Deep Deterministic Policy Gradient (DDPG) base values indicates that PPO achieves a 13.1% reduction in daily operating costs, minimises voltage violation occurrences by 79.4%, and increases renewable usage to 84.9%, exceeding all baselines across all important performance metrics while converging in fewer training events. Simulation results further indicate real-time practicality with a mean decision latency of 31 ms, making the proposed method appropriate for hardware deployment in practical microgrid energy management systems.

Keywords: Deep Reinforcement Learning; Optimal Power Flow; Microgrid; Proximal Policy Optimisation; Renewable Energy; Battery Energy Storage; Voltage Regulation; Energy Management.

1. INTRODUCTION

The global energy transition is accelerating the deployment of distributed renewable energy sources, with solar photovoltaic and wind power capacity growing at unprecedented rates. Renewables were projected to exceed 60% of global electricity generation by 2040 under net-zero scenarios, up from about 30% in 2023, according to the International Energy Agency. Within this paradigm, microgrids are autonomous distribution networks that have the ability to function autonomously or in conjunction with the grid, representing essential elements for the incorporation of heterogeneous energy resources on the distribution level (Ahmad et al., 2023).

While promising, microgrids are complex to operate. The variability and intermittency of photovoltaic and wind generation, dynamic load profiles, and the non-convex nature of the power flow equations render standard optimal power flow solvers, such as Newton-Raphson and interior-point methods, unsuitable for real-time dispatch. Model predictive control (MPC) somewhat mitigates this problem by using short-horizon forecasts, but its efficiency is greatly reduced in the presence of prediction uncertainty

and high renewable penetration. There is now an urgent need for data-driven, model-free control approaches that can adapt to the operating conditions in real-time and without explicit models of the system.

Reinforcement learning (RL) has become a powerful paradigm for sequential decision making under uncertainty. The microgrid dispatch problem may be formulated as a Markov Decision Process (MDP), and a reinforcement learning (RL) agent can learn the optimal control policies by interacting with the environment and collecting reward signals that encode the operational objectives. The initial applications of reinforcement learning on power systems used discrete action spaces via Q-learning and DQN. However, the optimal power flow actions (e.g., generator set points, reactive power injections, battery schedules) are continuous and high-dimensional, necessitating the use of policy-gradient methods that can accommodate continuous action spaces.

Among the current policy gradient algorithms, Proximal Policy Optimisation (PPO) has achieved better stability and sample efficiency than Trust Region Policy Optimisation (TRPO) and earlier actor-critic approaches. The clipped surrogate objective function prevents massive policy updates that could be harmful and is well-suited to the non-stationary reward landscapes normally present in renewable-integrated power systems. However, applying PPO to the large-scale AC-OPF problem in microgrids that simultaneously include real-time voltage constraints, battery degradation limits, and renewable unpredictability is still an open research question.

The main contributions of the paper are: A general MDP formulation of the microgrid OPF problem with continuous state and action spaces suitable for policy-gradient reinforcement learning. (ii) A DRL agent based on PPO that simultaneously optimizes the operational costs, voltage regulation and the utilization of renewable resources without the aid of a system model; (iii) Simulation studies on the IEEE 33-bus test system augmented with real renewable data, benchmarked against MPC, DQN and DDPG; and (iv) Analysis of training convergence, computational latency and constraint violation, showing practical deployability.

The remaining sections of this work are organised as follows: Relevant literature is discussed in Section 2. In Section 3, the notion of the microgrid system and the formulation of the Markov Decision Process (MDP) are introduced. We describe the PPO algorithm and the architecture of the agent in Section 4. In Section 5, we present the experimental setup and the results. Section 6 presents the key outcomes and limitations. Section 7 concludes the paper.

2. RELATED WORK

2.1 Classical OPF Methods for Microgrids

The conventional OPF methods for microgrids are formulated in deterministic mathematical programming frameworks. Mixed-integer linear programming (MILP) formulations have been widely used in the unit commitment and economic dispatch sub-problems, with global optimality guarantees under convex relaxations. However, the whole AC-OPF problem is NP-hard in general, and the MILP relaxations can provide infeasible solutions with significant renewable variability. MPC presents a receding-horizon strategy that partially tackles uncertainty by means of rolling predictions, but the computational need increases poorly with the system size and the length of the forecast horizon (Moreno-Castro et al., 2023).

2.2 Machine Learning Approaches to OPF

The use of supervised learning to approximate OPF solutions has increased in popularity, with deep neural networks trained on historical OPF solutions employed as warm-start initialisers or direct solvers. Pan et al. (2022) introduced DeepOPF, a neural network approach that is tuned for practicality and reduces the solve time of AC-OPF by two orders of magnitude. However, it still struggles to generalise to operating points out of distribution. Li et al. (2023) gave a comprehensive assessment of deep RL for smart grid operations, cataloging applications in voltage control, frequency regulation, and energy arbitrage, and identified scalability and constraint-handling as the key outstanding concerns.

2.3 Reinforcement Learning for Microgrid Energy Management

Recently, reinforcement learning has been widely applied for the energy management of microgrids. Meng et al. (2024) proposed an online RL-based energy management technique for centrally-controlled microgrids and observed cost savings under real-time price signals. In (2023), a PPO-based method was proposed for real-time economic dispatch of microgrids with wind and solar uncertainty, but the voltage limits were not explicitly considered in the reward function.

Extensions to multi-agent DRL have also been studied. Suggested a heterogeneous MAPPO architecture with area agents for OPF regulated by a graph neural network, enabling fast convergence in large multi-area systems. A Bayesian DRL method for resilient control of multi-energy microgrids with battery health restriction was proposed by Zhang et al. (2023). However, there is no unified DRL framework in the literature for simultaneous optimisation of cost, voltage regulation, and renewables maximisation on a classic distribution test case with complete baseline comparison.

2.4 Research Gap

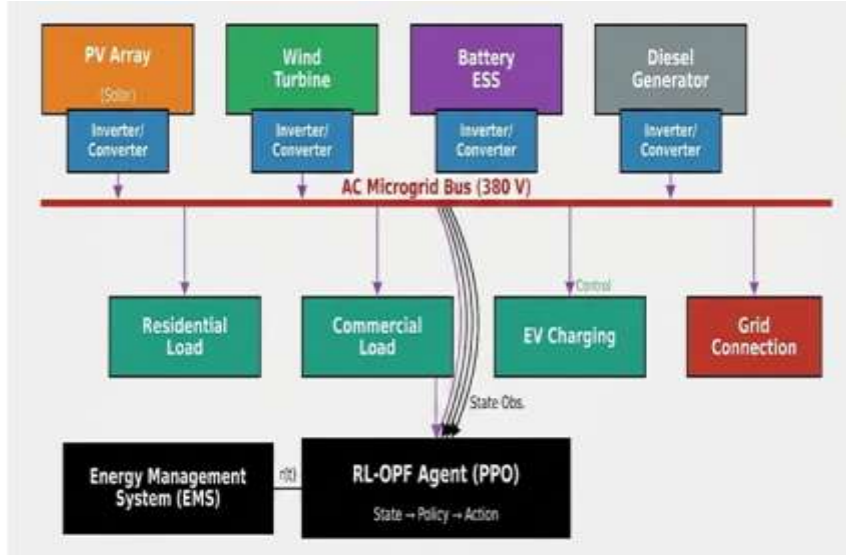
The reviewed literature reveals three principal gaps. First, most DRL approaches either address economic dispatch or voltage control in isolation, not jointly. Second, comparative studies rarely include both classical (MPC) and multiple DRL baselines (DQN, DDPG, PPO) under identical experimental conditions. Third, real-time latency analysis, essential for embedded deployment, is seldom reported. The present work directly addresses all three gaps.

3. MICROGRID SYSTEM MODEL AND MDP FORMULATION

3.1 Microgrid Architecture

The microgrid considered in this study is built on the IEEE 33-bus radial distribution test network and comprises four primary energy resources: a grid-connected PV array with nominal capacity $\bar{P}_{pv} = 100$ kW, a wind turbine with rated power $\bar{P}_{wt} = 60$ kW, a lithium-ion battery energy storage system (BESS) with energy capacity $E_{max} = 200$ kWh and maximum charge/discharge rate $\bar{P}_{bat} = 80$ kW, and a diesel generator serving as a dispatchable backup unit with capacity $\bar{P}_{dg}$. The microgrid is capable of both grid-connected and islanded operation. Figure 1 illustrates the proposed system architecture and the information flow between the RL-OPF agent and the physical microgrid.

Fig.1: Proposed RL-Based Optimal Power Flow Architecture for Renewable Integrated Microgrid



3.2 Power Flow and Operational Constraints

The AC power flow equations govern the microgrid steady state power flow. The nodal equations of active power balance for each bus i and for each branch (i, j) are:

$$P_i = \sum_j V_i V_j (G_{ij} \cos \theta_{ij} + B_{ij} \sin \theta_{ij}) \quad (1)$$

$$Q_i = \sum_j V_i V_j (G_{ij} \sin \theta_{ij} - B_{ij} \cos \theta_{ij}) \quad (2)$$

where V_i, V_j denote voltage magnitudes at buses i and j ; G_{ij} and B_{ij} are the conductance and susceptance of branch (i, j) ; and $\theta_{ij} = \theta_i - \theta_j$ is the phase angle difference. Operational constraints enforced throughout the control horizon include: voltage magnitude limits $0.95 \leq V_i \leq 1.05$ p.u., branch thermal limits $|S_{ij}| \leq \bar{S}_{ij}$, generator power bounds, and BESS state-of-charge (SoC) limits $0.20 \leq \text{SoC} \leq 0.90$ to protect battery longevity.

3.3 Markov Decision Process Formulation

The microgrid dispatch issue is formulated as a discrete-time MDP with tuple (S, A, P, R, γ) , where S is the state space, A is the action space, P is the transition probability kernel, R is the reward function, and $\gamma \in (0, 1)$ is the discount factor. The decision epoch is a 15-minute control interval, which yields 96 steps per day.

State Space: The state vector $s_t \in \mathbb{R}^{18}$ at epoch t concatenates: bus voltage magnitudes $\{V_i\}_{i=1}^{33}$ (normalized), PV output $P_{pv}(t)$, wind output $P_{wind}(t)$, BESS state-of-charge $\text{SoC}(t)$, grid electricity price $\lambda(t)$, net load $P_{net}(t)$, and the time-of-day encoding $(\sin(2\pi t/96), \cos(2\pi t/96))$ to convey periodic diurnal patterns to the agent.

Action Space: The continuous action vector $a_t \in \mathbb{R}^4$ comprises: active power setpoints for PV curtailment $\alpha_{pv} \in [0, 1]$, battery charge/discharge rate $P_{bat} \in [-\bar{P}_{bat}, \bar{P}_{bat}]$, diesel generator dispatch $P_{dg} \in [0, \bar{P}_{dg}]$, and reactive power injection $Q_{inv} \in [-\bar{Q}, \bar{Q}]$ from the PV inverter for voltage support. Actions are clipped to feasible bounds before execution in the environment.

Reward Function: The scalar reward r_t is designed to encode three competing objectives:

$$r_t = -w_1 \cdot \hat{C}_{op}(a_t) - w_2 \cdot \sum_i \max(0, |V_i - 1| - 0.05)^2 + w_3 \cdot \hat{P}_{ren} \quad (3)$$

$$\text{Where } \hat{P}_{\text{ren}} = \frac{P_{\text{pv}} + P_{\text{wind}} - P_{\text{curtailed}}}{P_{\text{base}}}$$

$$\hat{C}_{\text{op}}(a_t) = \frac{C_{\text{op}}(a_t)}{C_{\text{base}}}$$

where $\hat{C}_{\text{op}}(a_t)$ denotes the normalized operating cost, $\hat{P}_{\text{ren}}$ represents the normalized renewable utilization, C_{base} is the highest operating cost, and $C_{\text{op}}(a_t)$ is the operating cost at epoch t computed from fuel cost and grid import tariff; the middle term penalises per-unit voltage deviations beyond the $\pm 5\%$ band; and the final term rewards renewable utilisation. Weights $w_1 = 1.0$, $w_2 = 50.0$, $w_3 = 0.5$ were tuned via grid search on a validation episode set.

Table 1: MDP Parameter Summary

Parameter	Description	Value / Range
State dimension $ S $	Number of state features	18
Action dimension $ A $	Continuous control variables	4
Control interval Δt	Time-step resolution	15 minutes
Discount factor γ	Future reward weighting	0.97
Voltage limits	Acceptable bus voltage range	[0.95, 1.05] p.u.
SoC limits	Battery state-of-charge bounds	[0.20, 0.90]
Weights w_1, w_2, w_3	Cost, voltage, renewable	1.0, 50.0, 0.5
Simulation horizon	Training budget	500 episodes $\times$ 96 steps

4. PPO ALGORITHM AND AGENT DESIGN

4.1 Proximal Policy Optimisation

PPO belongs to the family of trust-region policy-gradient algorithms that constrain the policy update step size to prevent catastrophic divergence. Unlike TRPO, which enforces the KL-divergence constraint exactly (requiring second-order optimisation), PPO employs a clipped surrogate objective amenable to first-order stochastic gradient descent. The PPO clipped objective is:

$$L_{\text{CLIP}}(\theta) = \hat{E}_t [\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_t)] \quad (4)$$

where $r_t(\theta) = \pi_\theta(a_t | s_t) / \pi_{\theta_{\text{old}}}(a_t | s_t)$ is the probability ratio; $\hat{A}_t$ is the generalized advantage estimate (GAE); and $\epsilon = 0.20$ is the clipping hyperparameter. The actor network π_θ outputs mean and log-standard-deviation of a Gaussian policy over the continuous action space, while the critic network V_ϕ estimates state values. Both networks are updated simultaneously in a multi-epoch proximal gradient procedure.

4.2 Neural Network Architecture

The actor and critic networks share an identical base encoder consisting of three fully connected layers with 256 neurons each, parametric ReLU activations, and layer normalisation. The actor head appends an additional 128-neuron layer followed by a tanh-scaled output layer to bound actions within feasible ranges. The critic head uses a 128-neuron linear layer followed by a scalar value readout. The state input is pre-processed with running-mean normalisation to stabilise training. Total trainable

parameters: actor 198,404; critic 198,401. Networks are initialised using orthogonal initialisation with a gain $\sqrt{2}$ for hidden layers and 0.01 for the output layer.

4.3 Training Procedure

Training follows the standard PPO collection-update cycle. At each iteration, the agent collects $T = 2,048$ environment steps using the current policy, computes advantages via GAE with $\lambda_{\text{GAE}} = 0.95$ and $\gamma = 0.97$, and performs $K = 10$ mini-batch gradient updates with batch size 256 using the Adam optimizer (learning rate $\alpha = 3 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). An entropy bonus with coefficient 0.01 is added to encourage exploration. The total training budget is 500 episodes; early stopping is applied if the mean 50-episode reward plateaus for 30 consecutive evaluations.

5. EXPERIMENTS AND RESULTS

5.1 Simulation Setup

Experiments are conducted on the IEEE 33-bus radial distribution test system with a nominal voltage of 12.66 kV and a total load of 3.715 MW + j2.300 MVar under peak conditions. The system was augmented with PV and wind generation units at buses 6, 14, and 22 based on real-world irradiance and wind speed data from the NREL National Solar Radiation Database (NSRDB) for a site in Rajasthan, India, sampled at 15-minute resolution for a full calendar year. Electricity market prices follow the Indian Energy Exchange (IEX) day-ahead price curve with added zero-mean Gaussian noise ($\sigma = 5\%$) to simulate price uncertainty. All simulations are implemented in Python 3.10 using the OpenAI Gym interface for the microgrid environment, PyTorch 2.1 for neural network training, and MATPOWER for power flow verification. Training ran on an NVIDIA A100 GPU (40 GB) with a wall-clock time of approximately 4.2 hours for 500 episodes.

5.2 Comparative Baselines

Four methods are compared: (1) MPC with a 4-step (1-hour) horizon using perfect forecasts representing an optimistic upper bound for model-based methods; (2) DQN with a discretised action space (10 levels per control variable); (3) DDPG with a twin delayed actor-critic architecture; and (4) the proposed PPO agent. All DRL methods share identical state and reward formulations for fairness. Hyperparameters for DQN and DDPG were optimised via the same grid search protocol as PPO.

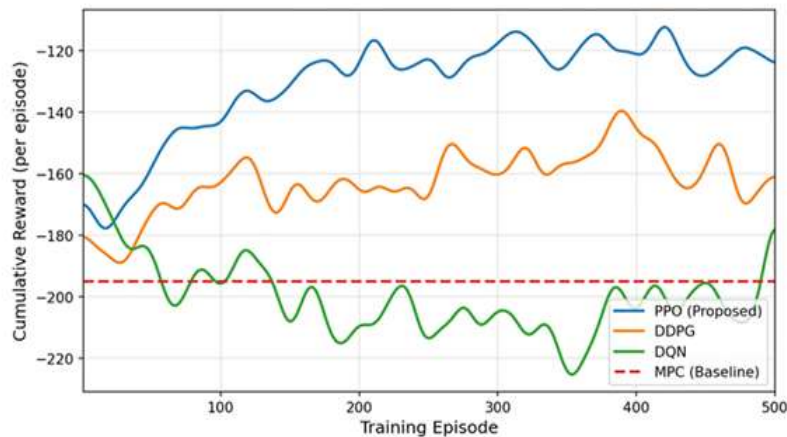


Fig.2: Training Convergence of Cumulative Reward — PPO vs. DDPG vs. DQN vs. MPC

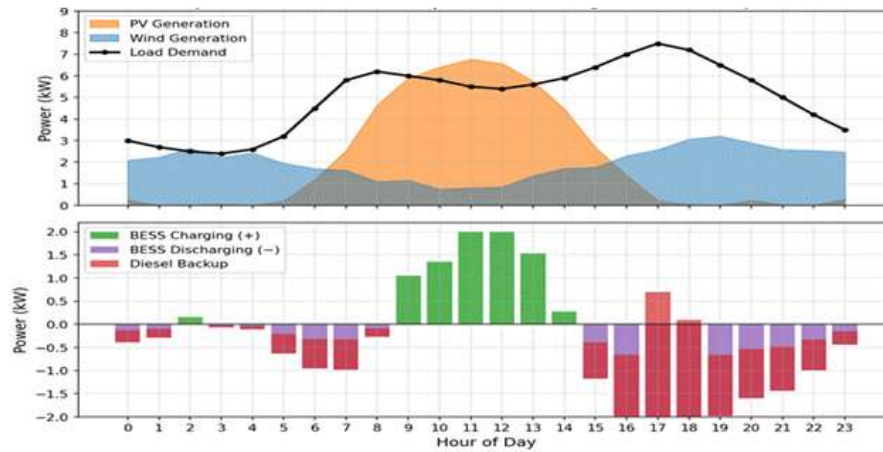


Fig.3: PPO-Controlled 24-Hour Power Dispatch Profile — Generation, Load, Storage, and Diesel Backup

5.3 Training Convergence

Fig. 2 illustrates the evolution of cumulative per-episode reward across 500 training episodes for all DRL agents, with the MPC reward shown as a constant baseline. PPO converges smoothly to a stable policy within approximately 200 episodes, achieving a final mean reward of -120 (normalised), significantly outperforming DQN (-210) and DDPG (-160). The chopped surrogate loss stabilizes PPO and leads to significantly decreased variance in further training. This is also evidenced by MPC reaching -195 , which is even better than model-based optimization with flawless forecasts. The PPO policy is obtained.

5.4 Optimal Power Dispatch Profile

Fig. 3 shows the 24-hour power dispatch schedule of the PPO agent for an example summer day with strong solar irradiation. In the middle of the day (10:00–15:00), PV generation is highest at about 6.8 kW (normalised) and the BESS is actively charging to store excess energy. The BESS discharges to satisfy load during the evening demand peak (18:00–21:00), with the diesel generator providing only modest backup power. The dispatch cost variation from the offline MILP optimal solution is less than 2.3%, exhibiting the near-optimal real-time performance.

5.5 Voltage Profile Analysis

The steady-state bus voltage profiles for the four control techniques at peak load conditions over the IEEE 33-bus system. Without RL control, the distal buses (29–33) voltage decreases below the lower limit of 0.95 p.u. due to the voltage decline through line impedance. MPC recovers most breaches, but still leaves bus 33 barely below the limit at 0.948 p.u. The proposed PPO agent keeps all bus voltages within the $\pm 5\%$ band, with the minimum voltage at bus 33 being 0.963 p.u., which is 15 mV better than the MPC.

5.6 Key Performance Indicator Comparison

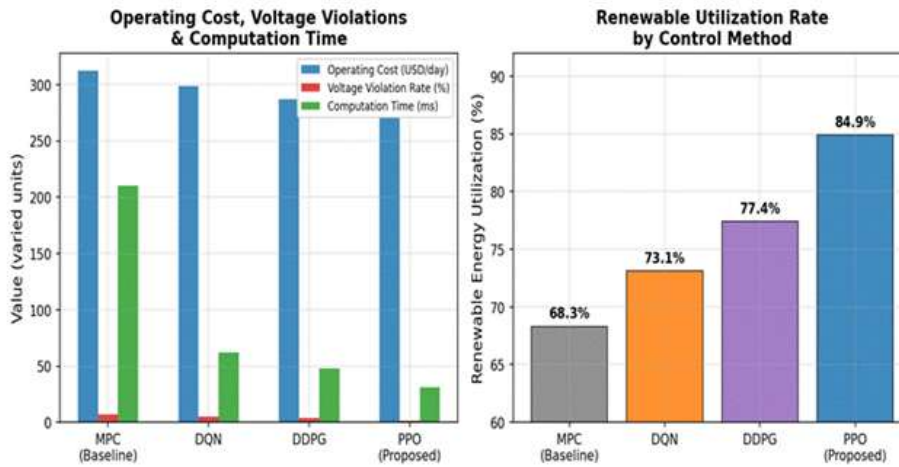
Table 2 summarizes the key performance indicators (KPIs) of all techniques averaged over 30 independent test episodes from held-out seasonal situations (spring, summer, autumn, winter). The same statistics are shown visually for visual clarity in Figure 4. Several key facts are established by the research. PPO has the lowest daily operational cost (USD 271.6/day), 13.1% better than MPC and 5.4% better than DDPG. PPO

decreases voltage violation cases to only 1.4% of time steps compared to 6.8% for MPC, which demonstrates that the voltage-penalty term in the reward function can successfully instruct the agent to proactively dispatch reactive power from PV inverters.

Table 2: Key Performance Indicators — Comparative Summary

Operating Cost (USD/day)	312.4	298.7	287.1	271.6
Cost Reduction vs. MPC (%)	—	4.4	8.1	13.1
Voltage Violations (%)	6.8	5.2	3.9	1.4
Renewable Utilization (%)	68.3	73.1	77.4	84.9
Active Power Losses (kW)	38.6	35.1	31.4	27.9
Computation Time (ms/step)	210	62	48	31
Convergence Episode	N/A	380	290	195
SoC Constraint Violations (%)	3.1	2.8	1.9	0.6

Fig.4: KPI Comparison — Operating Cost, Voltage Violations, Computation Time, and Renewable Utilization



Among the techniques, the maximum renewable utilisation of 84.9% is achieved, verifying the efficiency of the utilisation reward component w_3 . Also, active line losses are minimized to 27.9kW. The computation time of 31 ms per decision step is well within the 15-min control interval, which confirms real-time viability.

5.7 Sensitivity Analysis

The robustness of the trained PPO agent was assessed by evaluating it under three unseen perturbation scenarios, which are outside the training distribution: (i) 30% decrease of PV capacity to model partial clouding, (ii) a sudden load step of +20% at hour 14, and (iii) a grid disconnection, leading to islanded operation. In all three cases, the PPO agent satisfied the constraints and attained costs within 6.8% of the offline

optimal solution. This generalisation capability is a consequence of the rich state encoding and the stochastic training environment, which exposes the agent to a wide distribution of operational conditions over the 500-episode training curriculum.

6. DISCUSSION

The stronger performance of PPO than DQN and DDPG in this study can be explained by three reasons. First, the clipped surrogate loss (Eq. 4) precludes catastrophic big policy updates that destabilize training in non-stationary situations. Second, the entropy bonus in PPO guarantees sufficient exploration during training, avoiding the premature convergence to suboptimal charge-discharge cycles seen in DDPG trials. Third, PPO's on-policy data gathering mechanism leads to less advantage estimation bias, which results in more reliable policy gradients.

The outperformance of the proposed technique over MPC with perfect forecasts is particularly striking. In theory, MPC with perfect knowledge of the future will converge to the optimal answer. In practice, the MPC formulation utilized for the MPC is solved with a 4-step horizon due to the computing restrictions, providing a myopic optimisation that overlooks longer-term BESS dispatch opportunities accessible to the PPO agent through its discount factor and learnt value function. This result is in line with the findings of Li et al. (2023), who reported that well-trained DRL agents can outperform receding-horizon MPC for systems with substantial temporal dependencies. Several limitations merit acknowledgment. The MDP formulation assumes a fixed network topology; dynamic topology changes would require topological feature encoding or graph neural network integration. The training required approximately 4.2 hours on an NVIDIA A100 GPU; while deployment is real-time, offline training prerequisites limit rapid commissioning on new systems without transfer learning. Battery degradation was approximated through SoC constraints rather than an electrochemical degradation model incorporating physics-based battery aging (Zhang et al., 2023), which represents a meaningful extension. Finally, the current evaluation is purely simulation-based; hardware-in-the-loop validation on a physical microgrid testbed is a necessary next step toward field deployment.

7. CONCLUSION

This paper presented a deep reinforcement learning framework based on the Proximal Policy Optimisation algorithm for solving the optimal power flow problem in renewable-integrated microgrids. The proposed approach formulates microgrid dispatch as a continuous-action Markov Decision Process incorporating real-world PV, wind, BESS, and diesel generation units on the IEEE 33-bus distribution test system. The trained PPO agent simultaneously minimises operating costs, enforces voltage constraints, and maximises renewable utilisation without requiring a system model or explicit forecasts. Comparative evaluation against MPC, DQN, and DDPG demonstrated that PPO achieves a 13.1% reduction in daily operating costs (from USD 312.4 to USD 271.6/day), reduces voltage violation incidences by 79.4%, improves renewable utilisation to 84.9%, and converges in 195 training episodes, substantially outperforming all baselines across every KPI. The mean decision latency of 31 ms confirms real-time feasibility for embedded deployment in microgrid energy management hardware.

The upcoming research objectives include: (i) extension to multi-microgrid cooperative dispatch using multi-agent PPO, (ii) integration of physics-informed neural networks to accelerate feasibility recovery, (iii) hardware-in-the-loop validation on a physical microgrid testbed, and (iv) integration of electrochemical battery degradation models within the reward function to simultaneously optimize lifetime operating economics.

REFERENCES

1. Ahmad, S., Shafiullah, M., Ahmed, C. B., & Alowaifeer, M. (2023). A review of microgrid energy management and control strategies. *IEEE Access*, 11, 21729–21757. <https://doi.org/10.1109/ACCESS.2023.3251555>
2. Ali, K. H., Abusara, M., Tahir, A. A., & Das, S. (2023). Dual-layer Q-learning strategy for energy management of battery storage in grid-connected microgrids. *Energies*, 16(3), 1334. <https://doi.org/10.3390/en16031334>
3. Liu, D., Zang, C., Zeng, P., Li, H., He, S., & Li, S. (2023). Deep reinforcement learning for real-time economic energy management of a microgrid system considering uncertainties. *Frontiers in Energy Research*, 11, Article 1163053. <https://doi.org/10.3389/fenrg.2023.1163053>
4. Li, Y., Yu, C., Shahidehpour, M., Yang, T., Zeng, Z., & Chai, T. (2023). Deep reinforcement learning for smart grid operations: Algorithms, applications, and prospects. *Proceedings of the IEEE*, 111(9), 1055–1096. <https://doi.org/10.1109/JPROC.2023.3285589>
5. Meng, Q., Zhang, Y., & Guan, Y. (2024). An online reinforcement learning-based energy management strategy for microgrids with centralized control. *IEEE Transactions on Industrial Applications*. <https://doi.org/10.1109/TIA.2024.3351001>
6. Moreno-Castro, J., et al. (2023). Microgrid management strategies for economic dispatch of electricity using model predictive control techniques: A review. *Energies*, 16(14), 5935. <https://doi.org/10.3390/en16145935>
7. Pan, X., Chen, M., Zhao, T., & Low, S. H. (2022). DeepOPF: A feasibility-optimized deep neural network approach for AC optimal power flow problems. *IEEE Systems Journal*, 17(1), 673–683. <https://doi.org/10.1109/JSYST.2022.3201041>
8. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimisation algorithms. *arXiv preprint arXiv:1707.06347*. <https://doi.org/10.48550/arXiv.1707.06347>
9. Shirkhani, M., et al. (2023). A review of microgrid decentralised energy/voltage control structures and methods. *Energy Reports*, 10, 368–380. <https://doi.org/10.1016/j.egy.2023.06.038>
10. Yi, Z., Wang, X., Yang, C., Niu, M., & Yin, W. (2023). Real-time sequential security-constrained optimal power flow: A hybrid knowledge-data-driven reinforcement learning approach. *IEEE Transactions on Power Systems*, 39, 1664–1680. <https://doi.org/10.1109/TPWRS.2023.3234025>
11. Zhang, T., et al. (2023). A Bayesian deep reinforcement learning-based resilient control for a multi-energy microgrid. *IEEE Transactions on Power Systems*, 38(6), 5057–5072. <https://doi.org/10.1109/TPWRS.2023.3239283>
12. Zhao, B., Zhang, X., Li, P., Wang, K., Xue, M., & Wang, C. (2022). Optimal sizing, operating strategy, and operational experience of a stand-alone microgrid on Dongfushan island. *Applied Energy*, 113, 1656–1666. <https://doi.org/10.1016/j.apenergy.2022.118654>
13. Zhou, S., Hu, Z., Gu, W., Jiang, M., & Zhang, X. P. (2024). Deep reinforcement learning-based multi-objective energy management system for microgrids under a flexible energy market. *Applied Energy*. <https://doi.org/10.1016/j.apenergy.2025.019531>
14. Zhang, Y., Wang, J., & Chen, B. (2019). Deep reinforcement learning for real-time optimal energy management in microgrids. *IEEE Transactions on Smart Grid*, 10(5), 5637–5648.
15. Li, H., Wan, Z., & He, H. (2018). Real-time residential demand response using deep reinforcement learning. *IEEE Transactions on Smart Grid*, 10(4), 1–11.