Control of Small-Scale PV Parks for Grid Stability using Artificial Intelligence

Antonis P. Papadakis[1,2,] **Sofia Nikolaidou**[1], **Vasilis Hadjigeorgiou**[1], **Eleni Constantinide**[1]
[1]KYAMOS LTD, 37 Polyneikis Street, Strovolos, 2047, Nicosia, Cyprus
[2]Frederick University, 7. Y. Frederickou, Pallouriotisas, 1036, Nicosia, Cyprus

Abstract— Small-scale photovoltaic (PV) parks increasingly connect to low-voltage (LV) feeders, yet many installations lack SCADA connectivity, limiting real-time observability and controllability during overvoltage events. This paper proposes an asynchronous deep reinforcement learning (ADRL) framework for autonomous active-power control of PV inverters. A low-cost edge controller (Raspberry Pi) acquires inverter measurements and issues setpoints, while a Pandapower-based digital twin provides fast power-flow feedback for training on a GPU-enabled high-performance computing (HPC) cluster. An actor–critic policy is trained with the Asynchronous Advantage Actor–Critic (A3C) algorithm and a continuous Beta-distributed action representing an active-power dispatch fraction. The reward heavily penalizes voltage-limit violations while minimizing unnecessary curtailment. Simulation studies on a weak LV feeder indicate that the learned policy maintains voltage within limits and produces smoother, less conservative curtailment than a rule-based baseline, improving renewable utilization. The proposed architecture supports centralized training and decentralized low-latency execution, providing a practical pathway for scalable DER management in weak distribution grids.

Keywords—Grid stability; PV inverter control; asynchronous deep reinforcement learning; A3C; power curtailment; smart grids; high performance computing;

I. INTRODUCTION

High penetration of small-scale photovoltaic (PV) parks is increasingly challenging low-voltage (LV) distribution networks, especially in feeders with limited short-circuit capacity. During high irradiance and low demand, reverse power flow can cause local voltage rise, transformer and line loading, and nuisance inverter tripping. Effective mitigation requires fast inverter-level actions (e.g., active power curtailment and, where available, reactive power support), but many small PV parks lack Supervisory Control and Data Acquisition (SCADA) connectivity, limiting observability and controllability.

Distribution network operators therefore rely on static rule-based schemes (Volt–VAR and Volt–Watt curves) and conservative protection settings. While these methods are simple and locally robust, they are rarely optimal at feeder level and often curtail more energy than necessary because they do not account for changing grid strength, topology, and co-located generation and load.

In weak grids, stability depends not only on steady-state voltage but also on dynamic interactions between inverter controls and grid impedance. Reviews and measurement studies show that short-circuit power, frequency-dependent impedance, digital delays, and phase-locked loops can strongly influence inverter stability, and that simplified simulation tools may fail to capture such effects [1], [2]. These characteristics motivate data-driven control policies that can adapt to nonlinear, time-varying conditions.

Deep reinforcement learning (DRL) provides a model-free mechanism to learn control policies through interaction with a grid environment. Actor–critic methods are particularly attractive because they support continuous actions while maintaining stable learning through value-function baselines.

This paper presents an ADRL-based control framework for PV inverter active-power curtailment. A Raspberry Pi edge unit acquires inverter measurements and executes a trained policy locally, while training is performed asynchronously on a GPU-enabled HPC cluster using a Pandapower-based digital twin [3]. The learned policy is evaluated in a weak LV feeder and compared against a rule-based baseline, demonstrating smoother control actions and improved renewable utilization while respecting voltage limits.

The remainder of this paper is organized as follows. Section II reviews related work and introduces the actor–critic formulation. Section III describes the proposed architecture, training procedure, and data generation. Section IV presents simulation results. Section V concludes the paper and outlines future work.

II. LITERATURE REVIEW

A. *Reinforcement Learning*

Traditional approaches rely on deterministic, model-based controllers and predefined Volt–VAR or curtailment curves; however, these methods are often inadequate in weak grids due to nonlinear inverter–grid interactions, impedance variability, and rapidly changing operating conditions. Prior studies show that inverter stability is highly sensitive to short-circuit capacity, grid impedance characteristics, and internal control delays, making accurate modelling and robust control particularly challenging in low-voltage networks [1], [2], [4].

Recent advances in artificial intelligence have led to growing interest in reinforcement learning (RL) and deep reinforcement learning (DRL) for inverter-dominated power systems. Reviews report that learning-based controllers can outperform conventional proportional–integral, droop-based, and model-predictive control in adaptability and robustness, especially under uncertainty and nonstationary conditions [5], [4]. Actor–critic architectures (e.g., A2C/A3C, DDPG, TD3) and their multi-agent variants have been applied to Volt–VAR control, current regulation, and stability enhancement, improving voltage profiles and reducing losses while meeting grid-code constraints.

i) Actor-Critic Methods in Grid Applications

Actor-critic algorithms combine the benefits of policy-based and value-based reinforcement learning, making them particularly suitable for continuous control problems like inverter management. Recent research demonstrates the effectiveness of various actor-critic variants in power system applications. Ioannou et al. [6] evaluated multiple RL strategies for autonomous microgrid energy management, including Advantage Actor-Critic (A2C) alongside Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO). While DQN achieved superior performance with 73-95% cost reduction and near-zero state-of-charge imbalance across seasonal conditions, A2C demonstrated the synchronous policy-gradient approach's viability for real-time control with competitive computational efficiency. The Multi-Agent Actor-Critic (MAAC) framework has shown particular promise for coordinating multiple PV inverters. Rehman et al. [7] applied MAAC reinforcement learning to reactive power and voltage regulation in the IEEE-33 bus test system, achieving voltage controllable ratios of 0.6850 under decentralized control while maintaining grid voltage within ±5% limits and reducing voltage out-of-control ratios to 0.0275. This demonstrates the scalability of actor-critic methods to multi-agent coordination problems essential for distributed PV systems [8,9].

ii) Advanced Policy Gradient Approaches

Building on foundational actor-critic concepts, more sophisticated algorithms have emerged. Rajamallaiah et al. [10] implemented Twin Delayed Deep Deterministic Policy Gradient (TD3) control for three-phase grid-connected inverters with LCL filters, achieving 2.93% total harmonic distortion under nominal conditions with zero overshoot during transients. The TD3 approach, with its twin critics and delayed actor updates, demonstrated superior robustness to parameter mismatch, maintaining THD below 5% even with 50% inductance variation, outperforming both proportional-integral and model predictive control. For volt-var control applications, Beyer et al. [11] applied Deep Deterministic Policy Gradient (DDPG) to enable online learning in smart inverters using only local voltage measurements. The approach achieved voltage regulation to 1 ± 0.002 pu and reduced line voltage differences by up to 50% without requiring communication infrastructure, demonstrating the potential for decentralized learning with actor-critic methods.

These gaps highlight the need for scalable, learning-based control frameworks that combine centralized training with decentralized execution and are specifically tailored to the operational constraints of small PV installations in weak low-voltage networks. Recent related work further supports learning-based inverter and PV control, including deep reinforcement learning for transient stability improvement in grid-tied photovoltaics [12], comparative analyses of reinforcement learning versus neural-network controllers for inverter regulation [13], multi-objective DRL frameworks for adaptive power control in grid-forming inverters [14], learning Volt–VAR droop curves for coordinating PV smart inverters [15], and reinforcement-learning controllers augmented with short-term PV forecasts for voltage stability [16].

B. Asynchronous Deep Reinforcement Learning

The Actor-Critic algorithms integrate value-based and policy-based reinforcement learning methodologies within a unified framework, thereby leveraging the complementary strengths of both approaches. This hybrid architecture addresses fundamental limitations inherent to purely value-based or policy-based methods. The actor component employs a policy-based approach that naturally supports continuous action spaces, overcoming the discretization challenges that constrain traditional value-based methods such as Q-learning when applied to continuous control problems. Concurrently, the critic component provides value function estimates that serve to reduce the high variance characteristic of pure policy gradient methods, thereby stabilizing the training process and improving convergence reliability. This dual-network architecture enables the algorithm to optimize control policies for continuous photovoltaic dispatch factors while maintaining stable gradient estimates throughout the learning process, making it particularly well-suited for grid stability applications requiring precise, continuous actuation.

The Asynchronous Advantage Actor-Critic algorithm, widely known as A3C, represents a parallel reinforcement learning paradigm that leverages multiple asynchronous worker threads to accelerate policy learning while maintaining computational efficiency. In the context of power system control, asynchronous training across multi-GPU nodes accelerates convergence and enhances robustness against highly dynamic grid conditions. The distributed architecture enables parallel exploration of the state space, allowing multiple workers to simultaneously interact with independent instances of the power system environment. This parallelization not only reduces overall training time but also exposes the learning agent to a diverse range of operational scenarios concurrently, thereby improving the policy's generalization capabilities under varying grid conditions such as fluctuating photovoltaic generation, load variations, and topology changes. Unlike synchronous training methods that require all workers to complete their trajectories before updating shared parameters, A3C permits workers to compute and transmit gradients asynchronously, eliminating idle time and synchronization bottlenecks. This asynchronous paradigm is particularly advantageous for power system applications, where environmental dynamics exhibit high variability and computational resources must be utilized efficiently to achieve real-time control performance.

III. METHODOLOGY

A. *Model Design and Architecture*

The overall control architecture is organized in four steps:
(a) Edge sensing and local actuation.
Each PV park is equipped with a smart inverter and a local embedded controller (e.g., a Raspberry Pi 5). The controller continuously acquires measurements including local voltage magnitude, active and reactive power, and inverter status indicators (e.g., connection state and limits). At each control interval, the controller executes the trained policy and computes a continuous active-power dispatch factor u ∈ [0,1] (u = 1 indicates no curtailment). The resulting setpoint is sent to the inverter through a standard control interface (e.g., Modbus/SunSpec or vendor-specific APIs).
(b) State augmentation with a digital twin.
Real-time measurements are augmented by fast power-flow simulations using Pandapower [3]. This digital twin provides feeder-level indicators (e.g., bus voltages and transformer/line loading) and enables systematic generation of weak/strong grid operating scenarios for policy training and evaluation.
(c) Centralized asynchronous training.
Training is performed on a GPU-enabled HPC backend using the Asynchronous Advantage Actor–Critic (A3C) algorithm. Multiple worker processes interact with parallel simulation environments and asynchronously update a shared global actor–critic network, as illustrated in Figure 1.
(d) Decentralized deployment and online execution.
After training, the actor network is deployed to the edge controller, where inference runs locally with low latency and without requiring continuous communication. The backend can periodically retrain the policy using updated data and redeploy improved parameters.

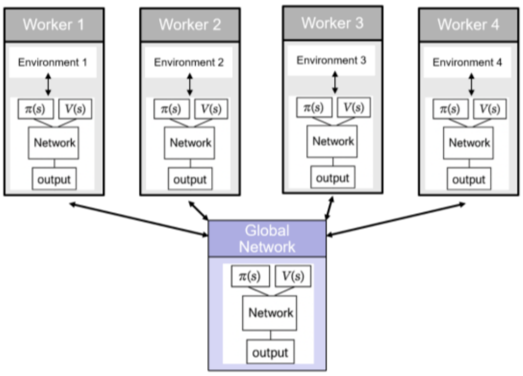

Figure 1: Proposed ADRL architecture with edge controllers, a Pandapower digital twin, and asynchronous A3C training.

At each discrete control time step, the agent observes a state vector constructed from local inverter measurements and Pandapower simulation outputs. Based on the observed state, the actor selects a continuous control action $u_t \in$ [0,1] representing the fraction of available PV power to inject. u_t= 1 indicates no curtailment or equivalently, an active-power curtailment command. The action is applied to the network model through a power-flow calculation, which determines the resulting grid state and the reward signal. The reward evaluates system performance with respect to voltage-limit compliance and minimization of unnecessary curtailment.

Each state transition, comprising the state, selected action, received reward, and subsequent state, is stored within the worker rollout buffer and used to compute policy and value gradients for A3C updates. During training, multiple asynchronous workers explore independent environment instances and update shared global parameters, enabling efficient learning under diverse operating conditions.

This architectural design separates the training and execution phases. Training leverages centralized information and HPC resources to learn control policies across diverse operating conditions. Execution runs on embedded edge controllers, where each PV park executes the policy using local measurements and a minimal set of global indicators. In the experiments reported in Section IV, the PV park is modeled as a single aggregated controllable unit (single agent) for clarity; the same framework extends to multiple parks through parameter sharing or multi-agent actor-critic training.

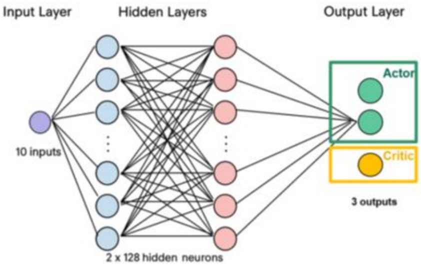

Figure 2: Actor–critic neural network used in A3C with separate actor and critic output heads.

B. *Training*

Each worker operates autonomously to collect rollout trajectories through asynchronous interaction with its dedicated environment instance. The rollout collection process involves three sequential stages executed independently by each worker thread. First, the worker samples actions from the current policy distribution based on observed grid states, determining appropriate control decisions for photovoltaic curtailment. Second, these sampled actions are applied to the worker's local power system environment, which executes the corresponding state transitions and computes the resulting grid conditions through power flow calculations. Third, the worker systematically stores all rollout data, including state observations, selected actions, received rewards, successor states, and auxiliary information such as policy log-probabilities and value estimates. This stored trajectory data serves as the basis for subsequent gradient computation and asynchronous parameter updates to the global network.

The rollout collection process employs a continuous action space parameterized by a Beta distribution. The policy network outputs are transformed using the Softplus activation function to obtain strictly positive parameters α and β, which define the Beta distribution from which actions are sampled: u ~ Beta(α,β). The critic network provides state value estimates through a neural network approximation: V(s) = NN(s).

Action selection proceeds by sampling from the parameterized Beta distribution, yielding a continuous dispatch factor u ∊ [0,1] (u = 1 indicates no curtailment). This action is subsequently applied to the power-flow simulation environment via Pandapower, where the PV output is reduced according to u, resulting in a new grid state characterized by updated voltage profiles and power flows.

Curtailment cost:

$$L_c(\mathrm{t}) = (1 - u_t)P_{avai}(t) \qquad (1)$$

, where P_{avail}(t) is the available PV power before curtailment and u_t ∈ [0,1] is the dispatched fraction (u_t = 1 indicates no curtailment).
The reward function is formulated to balance voltage regulation objectives against economic considerations. The curtailment cost is computed from the reduction in PV output, while voltage violations are assessed relative to the operating limits (V_{min}, V_{max}). The total reward is calculated as:

$$r(\mathrm{t}) = -\lambda\, L_V(\mathrm{t}) - \omega\, L_c(\mathrm{t})$$

(2)

$$L\_V(\mathrm{t}) = \Sigma_{\{\mathrm{b}\in\mathrm{B}\}}[\max(0, V_{\{b,t\}} - V_{max}) + \max(0, V_{min} - V_{\{b,t\}})\ ] \qquad (3)$$

In this study, V_{min} = 0.95 pu and V_{max} = 1.02 pu (adjustable per grid code), where λ >> ω, ensuring that voltage-limit violations dominate the objective relative to curtailment. This reflects the operational priority that voltage violations are unacceptable constraint breaches, whereas PV curtailment is a tolerable corrective action. All state transitions, actions, and rewards are stored as rollout data for subsequent gradient computation.

Moreover, for each temporal step within a worker's rollout trajectory, the Generalized Advantage Estimate (GAE) and return targets are computed locally using the following formulations in their simplified form.
The temporal-difference error, which serves as the foundation for critic updates, is defined as:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \qquad (4)$$

, where r_t represents the immediate reward, γ denotes the discount factor, and V(s) is the critic's value estimate.

The Generalized Advantage Estimate, utilized for policy gradient computation, incorporates exponentially-weighted temporal-difference errors:

$$A_t = \delta_t + \gamma\lambda A_{t+1} \qquad (5)$$

, where λ is the GAE parameter controlling the bias-variance trade-off between Monte Carlo and temporal-difference estimation.
The return target for critic training is computed as:

$$G_t = A_t + V(s_t) \qquad (6)$$

These quantities are computed locally within each worker process prior to gradient computation and synchronization.

Each worker thread independently computes gradients using the standard Asynchronous Advantage Actor-Critic loss formulation, comprising three components:
1. Policy Loss (Actor): The policy gradient loss maximizes the expected advantage-weighted log-probability of selected actions:

$$L_\pi = -\log \pi_t \cdot A_t \qquad (7)$$

2. Value Loss (Critic): The value function loss minimizes the mean squared error between predicted values and computed returns:

$$L_V = \left(V(s_t) - G_t\right)^2 \qquad (8)$$

3. Entropy Bonus: An entropy regularization term encourages exploration by penalizing overly deterministic policies.

Finally, the total worker loss is computed as the combined objective function of the three above losses.

Furthermore, the parameter synchronization procedure implements the standard Asynchronous Advantage Actor-Critic paradigm through a four-stage asynchronous update protocol. Each worker thread independently computes the gradients of the total loss function with respect to its local network parameters using the accumulated rollout data from the worker's trajectory. Computed gradients are then transmitted to the global network via an asynchronous push operation without synchronization barriers, allowing workers to operate independently and avoiding the computational overhead associated with synchronized batch updates. Upon receipt of worker gradients, the global network applies these gradients to update the shared parameters using the Adam optimization algorithm. Following the global update, the worker reloads the updated global parameters by copying them to its local network and resumes trajectory collection from its current environmental state using the refreshed policy and value function approximations. This asynchronous update mechanism enables parallel exploration across multiple workers while maintaining a single shared network, thereby improving sample efficiency and training stability compared to fully independent learning agents.

C. Data

1) Grid data

The proposed control framework targets weak low-voltage (LV) distribution feeders with high PV penetration. In general, the approach supports multiple small PV parks connected at different nodes along the feeder. For clarity, the case study used in this paper employs the simplified feeder shown in Figure 3 with a single representative aggregated PV park and an aggregated load, while preserving the same measurement-simulation-training workflow.

PV generation profiles at each park are derived either from real inverter measurements or from synthetic irradiance curves that capture typical clear-sky and partially cloudy conditions. Load profiles reflect realistic residential and commercial consumption patterns, including daily and seasonal variations. Voltage limits are imposed at all buses according to grid codes, and line and transformer ratings are explicitly modelled to capture thermal constraints and network losses.

This distribution network model serves as the environment for reinforcement learning (RL) agents during training and as the benchmark for evaluating the performance of the proposed controller. It allows the study of high PV penetration scenarios, including operating conditions where conventional rule-based control leads to voltage violations or excessive curtailment.

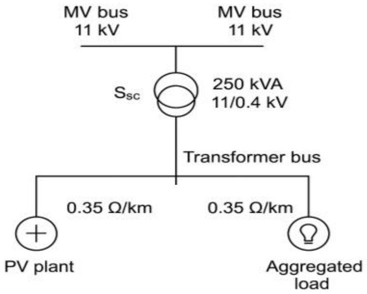

Figure 3: Simplified LV feeder model used as the Pandapower simulation environment.

2) Weather data

A synthetic irradiance-based model was developed to generate realistic photovoltaic power output profiles throughout diurnal cycles. The model employs a sinusoidal function to replicate the natural progression of solar irradiance, characterized by a gradual increase from sunrise, reaching maximum intensity at solar noon, followed by a symmetric decline toward sunset. During nighttime hours, photovoltaic generation is constrained to zero, reflecting the absence of incident solar radiation. This fundamental sinusoidal structure provides a physically consistent baseline representation of clear-sky solar resource availability.

To capture the stochastic variability inherent in real-world meteorological conditions, different noise patterns were superimposed on the baseline sinusoidal curve to simulate three distinct weather scenarios. Clear day conditions are characterized by high photovoltaic output with minimal variance, representing stable atmospheric conditions with unobstructed solar radiation. Partially cloudy conditions exhibit reduced mean generation levels punctuated by intermittent, short-duration power reductions corresponding to transient cloud cover obscuring solar panels. Cloudy day scenarios demonstrate persistently low generation levels with sustained high-frequency fluctuations, reflecting diffuse radiation conditions and continuous atmospheric attenuation. Regardless of the weather classification during daylight hours, photovoltaic output is uniformly set to zero during nighttime periods when solar radiation is unavailable. This multi-scenario approach ensures that the training dataset encompasses the full spectrum of generation variability encountered in operational distribution networks, thereby enhancing the robustness and generalization capability of the learned control policy.

3) Training data

The training dataset was generated through time-series power-flow simulations corresponding to approximately 266 day-equivalents of 5-minute operation when aggregating across all parallel environments. Four parallel environment instances were executed concurrently; each rollout consisted of 32 control timesteps at 5-minute resolution (160 minutes). Over the course of training, 600 asynchronous parameter updates were performed, yielding 4 x 32 x 600 = 76,800 agent-environment interactions. This dataset provides coverage of diverse operating conditions and facilitates robust policy learning across varying photovoltaic generation profiles and load demand patterns.

Two distinct grid strength scenarios were investigated to evaluate the control policy's adaptability to different network impedance characteristics. The weak grid scenario represents distribution networks with high feeder impedance relative to the short-circuit capacity at the point of common coupling. In such networks, even moderate photovoltaic power injections can induce significant voltage rise at low-voltage buses due to the predominantly resistive nature of voltage drops along distribution feeders. Conversely, the strong grid scenario characterizes networks with low feeder impedance or high short-circuit capacity, where the grid exhibits minimal voltage deviations in response to distributed generation variations. These contrasting scenarios enable comprehensive assessment of the learned control policy's performance across the spectrum of grid conditions encountered in practical distribution network operations.

The final dataset comprises features extracted from the Pandapower simulation environment [3], which collectively characterize the electrical behavior and operational state of the distribution network. The feature set encompasses short-circuit power at the point of common coupling, photovoltaic bus voltages, load power demand, transformer loading levels, and network impedance ratios. Additionally, the available active power capacity of each inverter, derived from the physics-based irradiance model, is incorporated to represent the potential generation prior to any curtailment actions. To capture temporal patterns and introduce continuity between daily operational cycles, time-based features including the day index and sinusoidal transformations of the time of day are included. These ten features collectively serve as inputs to the reinforcement learning model, comprehensively representing both the physical grid state and solar generation variability.

The PV bus voltage is included to reflect the local voltage magnitude at each generation connection point, enabling the detection of voltage rise phenomena associated with distributed generation. The load bus voltage provides visibility into voltage conditions at demand nodes, facilitating identification of overvoltage or undervoltage violations that may compromise power quality or equipment operation. Transformer loading indicates the utilization level of distribution transformers, revealing network stress and potential thermal or capacity constraints. Short-circuit power serves as an indicator of grid strength and fault current capability, which fundamentally influences voltage stability characteristics. Real-time load power informs the control policy of instantaneous demand conditions, while the available photovoltaic power before curtailment, extracted from the irradiance-based generation model, quantifies the potential renewable energy output. Finally, the impedance ratio captures the relationship between network resistance and reactance, which governs voltage drop behaviour and determines the effectiveness of reactive power control for voltage regulation.
Altogether, these 10 features are used as inputs to the proposed model, as they capture both the physical grid state and solar variability.

IV. RESULTS/DISCUSSION

The proposed controller is evaluated through time-series (quasi-static) power-flow simulations using the Pandapower digital twin and the simplified feeder of Figure 3. A rule-based baseline curtailment controller is used for comparison (labelled "teacher" in the plots). Performance is assessed using bus-voltage behavior, dispatched PV power, and control smoothness.

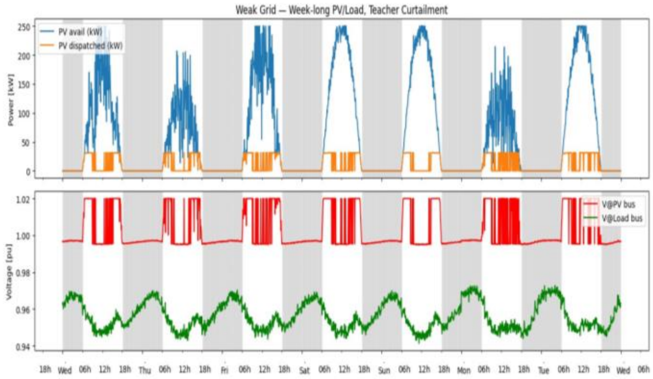

Figure 4: Weak-grid week-long profiles under a rule-based baseline controller (labelled "teacher" in the plot): available PV power, dispatched PV power, and bus voltages.

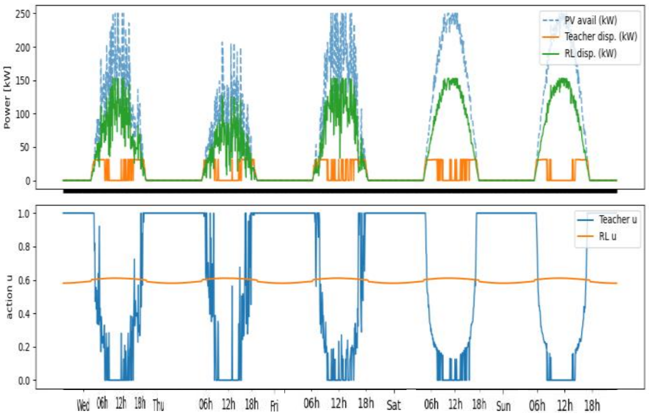

Figure 5: Comparison of ADRL and the rule-based baseline on a representative day in a weak grid: dispatched PV power and the dispatch action u (u = 1 indicates no curtailment).

Figure 4 illustrates that in weak-grid conditions, the baseline controller frequently curtails PV output during midday periods to keep the PV-bus voltage close to the upper limit (around 1.02 pu). The resulting dispatched power exhibits step-like behavior, indicating frequent control interventions. While this strategy maintains the voltage constraint, it can be overly conservative and may lead to unnecessary energy curtailment.

As shown in Figure 5, the ADRL policy produces smoother curtailment actions than the baseline and dispatches PV power closer to the available profile. In contrast, the baseline action exhibits rapid transitions between strong curtailment and no curtailment. The smoother ADRL behavior is desirable for practical deployment as it reduces setpoint chatter and can improve energy utilization while still respecting voltage constraints through the reward prioritization (λ >> ω).

Training remained stable with four parallel worker environments and 600 parameter updates (76,800 interactions), and the learned policy generalized across the considered irradiance patterns. Overall, these results indicate that actor–critic ADRL can provide a practical alternative to fixed Volt–Watt curtailment curves, particularly in feeders where grid strength and operating conditions vary over time. Field validation and explicit safety constraints (e.g., ramp-rate limits and communication delays) remain important next steps.

V. Conclusions/Future Work

This paper presented an asynchronous deep reinforcement learning (ADRL) framework for autonomous, voltage-aware PV inverter curtailment in small-scale LV installations that lack SCADA connectivity. The proposed architecture combines low-cost edge execution on a Raspberry Pi controller with centralized training on a GPU-enabled HPC backend using a Pandapower digital twin. Simulation results on a weak LV feeder show that the learned A3C policy can maintain voltages within limits while producing smoother and less conservative curtailment actions than a rule-based baseline.

Future work will focus on extending the action space to include reactive power support and coordinated multi-inverter control, integrating explicit safety constraints (e.g., ramp-rate limits, communication delays, and fail-safe fallback control), and validating the approach on larger feeders and field data.

Acknowledgment

This work is part of the **REALISATION-INVERTERS-PV** Project (Grant number: ENTERPRISES / ENERGY / 1123 / 0033), which is funded by the EU Recovery and Resilience Facility of the European Union - NextGenerationEU, and the Republic of Cyprus through the Research and Innovation Foundation under the Component 6.1 «REPowerEU» of the Cyprus Recovery and Resilience Plan. The ultimate objective of this project is to develop an autonomous system for the real-time control and monitoring of small-scale PV parks' inverters during periods of surplus supply or grid instability risks.

References

[1] Zhang Q, Mao M, Ke G, Zhou L, Xie B. Stability problems of PV inverter in weak grid: a review. IET Power Electronics. 2020; 13:2165-2174. https://doi.org/10.1049/iet-pel.2019.1049

[2] Höckel M, et al. Measurement of voltage instabilities caused by inverters in weak grids. CIRED. 2017; 2017(1):770-774. doi: 10.1049/oap-cired.2017.0997

[3] Thurner L, Scheidler A, Schäfer F, Menke JH, Dollichon J, Meier F, Meinecke S, Braun M. pandapower—An Open Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems. IEEE Transactions on Power Systems. 2018; 33(6):6510-6511.

[4] Massaoudi MS, Abu-Rub H, Ghrayeb A. Navigating the Landscape of Deep Reinforcement Learning for Power System Stability Control: A Review. IEEE Access. 2023; 11:134298-134317. doi: 10.1109/ACCESS.2023.3337118

[5] Kurukuru VSB, Haque A, Khan MA, Sahoo S, Malik A, Blaabjerg F. A Review on Artificial Intelligence Applications for Grid-Connected Solar Photovoltaic Systems. Energies. 2021; 14(15):4690. https://doi.org/10.3390/en14154690

[6] Ioannou I, Javaid S, Tan Y, Vassiliou V. Autonomous Reinforcement Learning for Intelligent and Sustainable Autonomous Microgrid Energy Management. Electronics. 2025; 14(13):2691. https://doi.org/10.3390/electronics14132691

[7] Rehman Au, Ali M, Iqbal S, Shafiq A, Ullah N, Otaibi SA. Artificial Intelligence-Based Control and Coordination of Multiple PV Inverters for Reactive Power/Voltage Control of Power Distribution Networks. Energies. 2022; 15(17):6297. https://doi.org/10.3390/en15176297

[8] Liu H, Wu W. Online Multi-Agent Reinforcement Learning for Decentralized Inverter-Based Volt-VAR Control. IEEE Transactions on Smart Grid. 2021; 12(4):2980-2990. doi: 10.1109/TSG.2021.3060027

[9] Guo G, Zhang M, Gong Y, Xu Q. Safe multi-agent deep reinforcement learning for real-time decentralized control of inverter based renewable energy resources considering communication delay. Applied Energy. 2023; 349:121648. doi: 10.1016/j.apenergy.2023.121648

[10] Rajamallaiah A, Karri SPK, Alghaythi ML, Alshammari MS. Deep Reinforcement Learning Based Control of a Grid Connected Inverter With LCL-Filter for Renewable Solar Applications. IEEE Access.

2024; 12:22278-22295. doi: 10.1109/ACCESS.2024.3364058

[11] Beyer K, Beckmann R, Geißendörfer S, von Maydell K, Agert C. Adaptive Online-Learning Volt-Var Control for Smart Inverters Using Deep Reinforcement Learning. Energies. 2021; 14(7):1991. https://doi.org/10.3390/en14071991

[12] Dewantoro G, Swain A, Patel N. Transient Stability Improvement of Grid-Tied Photovoltaics using Deep Reinforcement Learning. 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN). 2024:1-6. doi: 10.1109/INDIN58382.2024.10774216

[13] Abdelwahab SAM, Khairy HE, Yousef H, et al. Comparative analysis of reinforcement learning and artificial neural networks for inverter control in improving the performance of grid-connected photovoltaic systems. Scientific Reports. 2025; 15:24477. https://doi.org/10.1038/s41598-025-09507-9

[14] Rajak MK, Pudur R. Deep reinforcement learning framework for adaptive power control in grid-forming inverters: A multi-objective optimization approach. Journal of Renewable and Sustainable Energy. 2025; 17(2):026301. https://doi.org/10.1063/5.0249385

[15] Glover D, Dubey A. Learning Volt-VAR Droop Curves to Optimally Coordinate Photovoltaic (PV) Smart Inverters. IEEE Transactions on Industry Applications. 2025; 61(1):859-871. doi: 10.1109/TIA.2024.3472655

[16] Zalavadiya HB. Development of controller using Reinforcement learning with Short term PV forecast for grid voltage stability. Master's thesis, University of Siegen; 2025. https://elib.dlr.de/214743/