International Journal of Smart Grid (IJSmartGrid)
https://www.ijsmartgrid.org/ Online ISSN: 2602-439X
R. Mihramane et al., Vol. 10, No. 2, June, 2026 Research Article

# CTDE Multi-Agent Deep Q-Network for Grid-Interactive Energy Management of a Saharan Microgrid

Redouane Mihramane *‡, Sidi Salah Ech-Charqaouy*, Abdelkader Boulezhar*, Nizar Ech-Charqaouy*, Amjad Ech-Charqaouy*

*Department of Physical Sciences, Faculty of Sciences Ain Chock (FSAC), Hassan II University of Casablanca, Casablanca, Morocco
(redmihramane@gmail.com, echcharqaouy@gmail.com, boulezhar@yahoo.fr, nizarechcharqaouy@gmail.com, echcharqaouyamjad@gmail.com)

‡Corresponding Author: Redouane Mihramane, Department of Physical Sciences, Faculty of Sciences Ain Chock (FSAC), Hassan II University of Casablanca, Casablanca, Morocco. E-mail: redmihramane@gmail.com Tel.: +212 661910758

*Received: 25.07.2025 Accepted: 23.04.2026*

**Abstract-** We propose a grid-interactive, decentralized multi-agent Deep Q-Network (MA-DQN) with Centralized Training, Decentralized Execution for energy management in a Saharan hybrid microgrid (Boujdour, Morocco). Independent agents supervise PV, wind, diesel and batteries; a low-overhead PCC coordinator enforces import/export exclusivity, zero-crossing deadband and dwell time. A multi-objective reward balances cost, $CO_2$ and grid exchange, with penalties on ±10% voltage and SOC violations. On a 24-h case (demand 991.1 kWh), the learned policy delivers renewables-first dispatch (PV 56.4%, diesel 34.4%, grid 9.2%), battery throughput ±7.6 kWh, and 272.5 kg $CO_2$ (~275 g/kWh). Versus a PSO baseline, cost falls 1250→1120 MAD (−10.4%), $CO_2$ 318.9→272.5 kg (−14.6%), and inference latency 12.8→2.1 s (−83.6%). Sensitivity on the $CO_2$ weight β yields tunable cost–carbon trade-offs (β 0.1→2.0: 310→240 kg $CO_2$ at +15.5% cost). Stress tests (PV depression/outage, +20% evening peak) maintain supply and power-quality; ablation confirms reduced PCC chattering versus single-agent DQN.

**Keywords:** Microgrid energy management, reinforcement learning, multi-agent DQN, CTDE, demand response

## 1. Introduction

The global energy transition is reshaping power systems under the twin imperatives of cutting greenhouse-gas emissions while preserving reliability and affordability in data-scarce, operationally volatile contexts [1], [2]. Hybrid renewable architectures - especially PV–wind portfolios complemented by storage and dispatchable backup - have expanded rapidly; yet intermittency, forecasting uncertainty, and network constraints still complicate real-time balancing and long-term planning [3], [4], [5].

### *1.1. PV and Carbon Background*

Rapid cost declines and deployment have made PV a central lever for decarbonization, but its net climate benefit ultimately depends on system-level operation that curtails spillage, limits diesel dependence, and avoids unnecessary imports from carbon-intensive grids, [2], [5], [6]. Recent assessments stress that renewable build-out must be coupled with operational strategies that internalize $CO_2$ impacts and policy priorities, rather than assuming carbon neutrality by construction [7], [8].

### *1.2. Saharan Microgrid Context*

In Saharan regions such as Boujdour (southern Morocco), extreme irradiance, reliable nocturnal wind, harsh thermal conditions, and limited grid robustness create a compelling yet challenging setting for hybrid microgrids [9]-[11]. Typical distribution feeders are long, radial low-voltage (LV) lines with non-negligible impedance and sparse instrumentation; operational feasibility therefore hinges on coordinated use of PV, wind, batteries, a diesel genset, and a bidirectional point of common coupling (PCC) to the main grid, while respecting ±10% LV voltage envelopes and battery state-of-charge (SOC) safety constraints [12]-[15]. These characteristics motivate supervisory controllers that are both network-aware and policy-tunable (cost, carbon, autonomy).

**Cite this article as:** R. Mihramane, S.S. Ech-Charqaouy, A. Boulezhar, N. Ech-Charqaouy, and A. Ech-Charqaouy, "CTDE Multi-Agent Deep Q-Network for Grid-Interactive Energy Management of a Saharan Microgrid" *International Journal of Smart Grid (ijSmartGrid)*, Vol. 10, No. 2, pp. 120-140, June, 2026.

### *1.3. Limits of Conventional Supervisory Control*

Deterministic scheduling (e.g., day-ahead mixed-integer linear programming, MILP), rule-based heuristics, and meta-heuristics such as particle swarm optimization/genetic algorithms (PSO/GA) have produced valuable baselines; however, they are brittle under regime shifts (seasonality, demand surges, partial outages) and costly to re-solve in real time. They also struggle to embed discrete field actuation and practical PCC rules - import/export exclusivity, zero-crossing deadband, and dwell-time to prevent chattering - while enforcing voltage feasibility along LV feeders at runtime [6], [16]. This motivates learning-based controllers that generalize across operating conditions, ingest local and broadcast signals, and encode grid-feasible actions within the policy.

### *1.4. From RL to DRL in Microgrids*

Reinforcement learning (RL) optimizes sequential decisions through interaction with an environment and has shown promise for distributed energy management in islanded and grid-connected settings [17]-[22]. Deep RL (DRL) scales RL to high-dimensional, nonlinear dynamics typical of hybrid microgrids with uncertain exogenous inputs and network constraints. Within DRL, the Deep Q-Network (DQN) is particularly suited to discrete actuation - diesel on/off, battery charge/hold/discharge, and tiered import/export at the PCC, offering stable training and edge-friendly inference; continuous-action methods (DDPG/SAC/PPO) remain complementary for fine set-point refinement and are considered as future extensions [23]-[27].

### *1.5. State of the Art and Interaction with the Main Grid*

A growing literature tackles centralized energy management with network limits and cost signals, as well as single- and multi-agent DRL for distributed coordination [17], [18], [20], [26]-[28], alongside broader reviews of hybrid renewable systems and operational challenges [3]-[5]. Yet comparatively fewer studies incorporate explicit PCC rules (exclusivity deadband, dwell-time) together with voltage-aware penalties on long LV feeders, while quantifying cost–carbon–autonomy trade-offs under Saharan conditions and benchmarking against strong offline baselines [27]-[29].

### *1.6. Research Gap*

There remains a need for a decentralized, grid-interactive, and voltage-aware DRL controller that: (i) coordinates heterogeneous sources under centralized training and decentralized execution (CTDE); (ii) enforces realistic PCC policies and action feasibility with discrete field commands; (iii) treats carbon as a policy parameter rather than a fixed price; and (iv) reports seasonally robust performance with transparent benchmarks and sensitivity to policy weights [6]-[8], [27], [29].

### *1.7. Objective*

To design and validate such a controller on a realistic Boujdour digital twin (31-node LV feeder, Morocco) with summer/winter scenarios and explicit main-grid interaction.

### *1.8. Contributions*

This work delivers:(i) a CTDE multi-agent DQN with independent learners for PV, wind, diesel, and batteries, plus a lightweight PCC coordination policy enforcing import/export exclusivity, zero-crossing deadband, and dwell-time (discrete actuation well-matched to field hardware) [17]-[20]; (ii) a network-feasible reward that jointly optimizes monetary cost, $CO_2$, and grid exchange while penalizing voltage (±10%) and SOC violations, aligning with LV practice and battery safety [11]-[15]; (iii) a seasonal digital-twin benchmark of a Saharan hybrid microgrid with explicit main-grid interaction, yielding a renewables-first learned dispatch and, versus a PSO baseline, cost −10.4%, $CO_2$ −14.6%, and inference latency −83.6%, all without voltage/SOC violations, in line with statistical parametric simulation studies grounded in real-world data [30]; (iv) a policy-weight sensitivity that treats the carbon term as a policy knob, mapping cost–carbon trade-offs across β sweeps [6]-[8]; (v) a robustness evaluation under unseen stressors (PV depression, partial PV outage, +20% evening peak) and an ablation showing that the multi-agent design reduces PCC chattering and smooths grid interaction relative to a single-agent DQN; (vi) a deployment pathway arguing the edge-friendliness of discrete DQN and outlining continuous-action DRL (DDPG/SAC/PPO) as a future refinement for set-point smoothing without changing the supervisory logic.

## 2. Related Works

Research on microgrid energy management systems (EMS) has evolved from deterministic scheduling and rule-based control toward data-driven and learning-based supervision capable of handling uncertainty, multi-objective trade-offs, and grid–microgrid interactions [2], [6], [16], [31]. In renewable-dominated and weakly interconnected settings - typical of Saharan sites - this evolution is driven by intermittent resources, long low-voltage feeders, and stringent power-quality constraints [3], [4], [7], [8], [10].

### *2.1. Deterministic, Heuristic, and Metaheuristic EMS*

Conventional EMS formulations cast scheduling as mathematical programming - e.g., LP, MILP, NLP, or DP - aimed at minimizing operating cost under explicit constraints and prespecified forecast scenarios [3], [4]. While such models afford optimality guarantees when forecasts and plant models are dependable, their performance degrades under regime shifts and model–plant mismatch, conditions that are commonplace in desert microgrids with pronounced diurnal and seasonal variability [2], [7], [8]. Heuristic controllers (rule-based and expert systems) improved practicality by codifying operator know-how (e.g., diesel-priority rules, SOC thresholds), yet their fixed rules lack the adaptivity required for hour-to-hour fluctuations and contingencies [4], [11], [17]. Metaheuristics (PSO, GA, ACO) broaden the search to tackle nonlinear, high-dimensional, multi-objective formulations in planning and quasi-static scheduling [19], [32]. They remain valuable baselines, but latency and convergence sensitivities limit their suitability for real-time dispatch when operating conditions evolve faster than optimization loops can track reliably [23], [33]. Recent surveys on hybrid-renewable optimization echo this divide - strong offline optimization

versus weaker online adaptivity - thereby motivating learning-based supervisory control [2], [4], [6]. In parallel, DC-microgrid architectures have demonstrated cost-effective pathways to bolster supply resilience in grid-connected communities [34].

### *2.2. Reinforcement Learning and Deep RL for Microgrid Control*

Reinforcement Learning (RL) learns sequential decision policies by interacting with the environment and receiving scalar rewards, removing the need for fully specified models and enabling adaptation to stochastic conditions [17], [18], [20], [28], [29]. Early Q-learning studies demonstrated fuel savings and improved storage utilization in hybrid microgrids under variable renewables and loads [9], with further applications to demand flexibility and residential contexts [19], [35].

Deep RL (DRL) scales RL to high-dimensional states via neural function approximation, stabilizing training and improving generalization under uncertainty [13], [36]. The Deep Q-Network (DQN) - with experience replay and target networks - has been widely adopted for discrete supervisory actions (diesel on/off, battery charge/hold/discharge modes, tiered PCC exchanges), aligning well with common EMS actuation granularity [23], [37], [38]. Continuous-action methods (DDPG, SAC, PPO) target fine set-point control and have reported gains in renewable penetration and $CO_2$ reduction in suitable settings, but often at increased implementation and tuning complexity [27], [28]. Hybrid approaches (e.g., DRL with fuzzy logic) also appear in rapidly varying systems to mitigate nonlinearity and measurement noise [9].

For weakly interconnected Saharan microgrids - characterized by costly diesel logistics, unstable grid back-up, and harsh ambient conditions - DRL enables cost- and carbon-aware dispatch while enforcing feasibility via SOC envelopes and voltage limits [11], [24]. Recent contributions in grid-edge control and voltage regulation using dueling and double-Q variants further underline the relevance of value-based DRL when network constraints matter [24], [25], [37].

### *2.3. Multi-Agent DRL and Grid-Interactive, Voltage-Aware EMS*

As microgrids scale in heterogeneity (PV, wind, diesel, multiple batteries, controllable PCC), multi-agent formulations become natural. Centralized-training, decentralized-execution **(**CTDE**)** designs combine global learning signals with independent learners per resource to reduce the joint action complexity while retaining implementable local policies[27], [29], [36]. Within this paradigm, coordination at the PCC - import/export exclusivity, anti-chattering deadbands with dwell time, and tariff-aware exchange - can be encoded either through shared constraints or a lightweight coordinator policy without sacrificing decentralization at execution time [11]-[13], [24].

Voltage-aware DRL is particularly pertinent in long LV radial feeders where impedance-driven drops can breach ±10% limits during high PV export or abrupt load ramps. Recent grid-interactive DRL work addresses these issues with value-based methods (dueling DDQN) and network-informed rewards/penalties that internalize nodal voltages and branch flows during policy learning [24], [37]. Beyond performance, robustness and security are receiving growing attention in EMS design across Industry-5.0 cyber-physical contexts, reinforcing the case for interpretable, network-feasible control logic [6], [22].

### *2.4. Positioning of the Present Research*

Against this backdrop, the present study advances the state of the art along four axes:

CTDE multi-agent DQN with independent learners (PV, wind, diesel, batteries) and PCC-aware coordination that enforces import/export exclusivity and anti-chattering, matching the discrete actuation structure of practical EMS [11]-[13], [24], [29], [37].

A multi-objective formulation balancing cost**,** $CO_2$**,** import reliance**,** and network quality (voltage/SOC), with policy-parameter sensitivities instead of assuming a fixed carbon price, consistent with emerging decarbonization analyses and policy diversity [2], [5]-[8].

A realistic digital twin of a Saharan hybrid microgrid (Boujdour) used for training and evaluation under seasonal scenarios and unseen stresses (PV derating, demand surges, battery/PCC outages), reflecting field constraints documented in prior work [9]-[11], [13].

A comparative evaluation against a PSO baseline and a single-agent DQN ablation to quantify gains in cost, emissions, reliability, and voltage compliance, bridging the gap between planning-oriented metaheuristics and real-time learning-based EMS [16], [19], [20], [23], [27].

## 3. Microgrid Description and Operating Environment

### *3.1. Site and Grid Context*

The studied microgrid is located in Boujdour, a coastal town in the Moroccan Sahara (26.13° N, 14.32° W). The site couples high solar irradiance with persistent wind regimes - particularly nocturnal winds - under relatively mild ambient temperatures, providing favorable conditions for hybrid, decentralized generation. Conversely, geographic remoteness and a quasi-linear radial low-voltage (LV) feeder impose stringent requirements on voltage regulation, service continuity, and cost-effective dispatch [7]-[13].

The LV distribution network supplies 218 households, two commercial facilities, and **two** industrial consumers (with peak demands of 15 kW and 20 kW), and extends 2 170 m in a radial topology. The microgrid can operate in islanded or grid-connected mode via an MV/LV substation equipped with a 250 kVA transformer, automated switchgear, and a bidirectional advanced metering interface to the national ONEE grid. This interconnection enables dynamic import/export in response to local generation and tariff signals.

Asset Portfolio and Electrical Interface

The Boujdour microgrid is built around a 31-node radial low-voltage (LV) feeder interconnected with the national grid via a bidirectional point of common coupling (PCC). Local

generation and storage include three PV arrays (44 kW, 10 kW, and 6 kW), a 90 kW wind turbine, a 50 kVA diesel genset, and three lead-acid battery banks co-located with the PV plants. Direct PV–battery coupling enables localized charge/discharge, reduces feeder losses, and enhances resilience to generation or network contingencies. A supervisory layer of smart meters and controllers monitors active-power flows, bus voltages, and battery state-of-charge (SOC) to support coordinated control across the LV network. This architecture is consistent with well-established solar-hybrid system practices [39].

PCC operation and rules

The PCC at the MV/LV substation allows seamless islanded or grid-connected modes. Imports bridge renewable deficits; exports evacuate surplus, limiting diesel use. To stabilize grid interaction, we enforce: (i) import/export exclusivity (never both at once); (ii) a zero-crossing deadband around PG=0; and (iii) a dwell-time before any sign inversion (anti-chattering). PCC set-points are tiered {−50,−25,0,+25,+50} kW to match field actuation.

*(a) Evidence (24 h test)*

With Δt=15 min, deadband ζ=5 kW, and the above tiers, the controller (vs. a tiers-only baseline) blocked 13 sign-change attempts, increased neutral intervals by +13 steps (18→31), and reduced gross import 375.0→350.0 kWh (−6.7%) and gross export 243.75→181.25 kWh (−25.6%). This reflects fewer micro-oscillations around zero and less superfluous energy shuttling, while preserving import during deficits and export during surplus.

*(b) Network feasibility*

Voltages are in p.u. with an LV compliance band ±10% at all monitored buses. Batteries respect SOC bounds [$SOC_{min}$,$SOC_{max}$] and power/ramp limits. At each step a fast power-flow screen masks any action that would violate SOC, device limits, or the LV voltage band, ensuring feasible-by-construction decisions.

*(c) Data granularity*

Demand and renewable inputs are processed at hourly resolution using summer/winter scenarios aligned with the 24-h horizon and discrete actuation.

Single-line diagram. Fig. 1. shows the LV network and asset interconnections (PV–wind–diesel–BESS–PCC). The full technical specification table is provided in Supplementary S1.

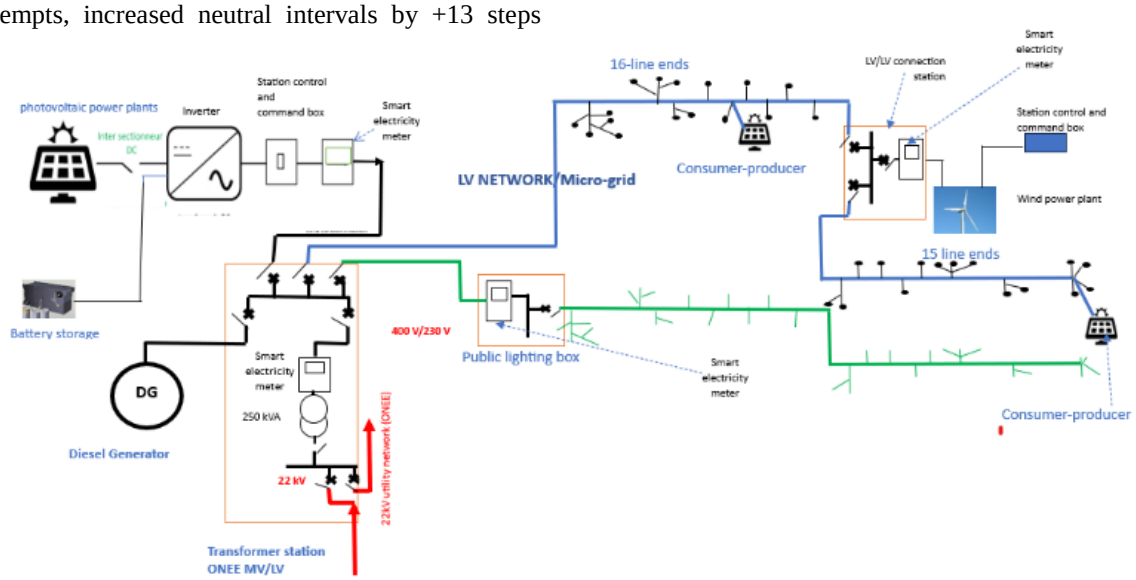



**Fig. 1.** Electrical single-line diagram of the studied microgrid integrating photovoltaic systems, wind generation, diesel generator, battery storage, and grid interconnection.

*3.2. Climatic Conditions and Renewable Resources*

*a) Wind*

ERA5T reanalysis indicates average speeds of ≈ 14 m/s in January and ≈ 17 m/s in July at 80 m hub height, with north–northeast winds dominating > 60 % of the time at speeds exceeding 20 km/h**.**

*b) Solar*

Year-round high insolation with mild temperatures (≈ 17 °C in January, ≈ 22 °C in July) favors PV conversion efficiency by limiting thermal losses.

These stable yet hourly intermittent resources provide a robust basis for hybrid operation, while motivating intelligent scheduling and storage control to accommodate short-term fluctuations [7], [8].

*3.3. Aggregate Load Profile*

The aggregated demand exhibits pronounced diurnal variability and seasonal shifts [40]. Off-peak nighttime load (02:00–06:00) drops to approximately 9.2 kW. Two distinct peaks are observed:

A morning peak around 10:00, reaching ≈ 64.2 kW, driven mainly by commercial and light industrial activity. A

dominant evening peak between 20:00–22:00, peaking at ≈ 110 kW, largely residential (lighting, cooking, entertainment).

Fig. 2. illustrates the typical hourly demand profile, highlighting the sharp evening peak and low nighttime consumption. The evening peak coincides with minimal PV output, increasing reliance on battery discharge and, if storage is insufficient, on the diesel generator. Mid-day surplus PV can be stored or exported via the PCC. This temporal mismatch between generation and demand motivates adaptive EMS strategies capable of load shifting and storage scheduling in real time [25], [41].

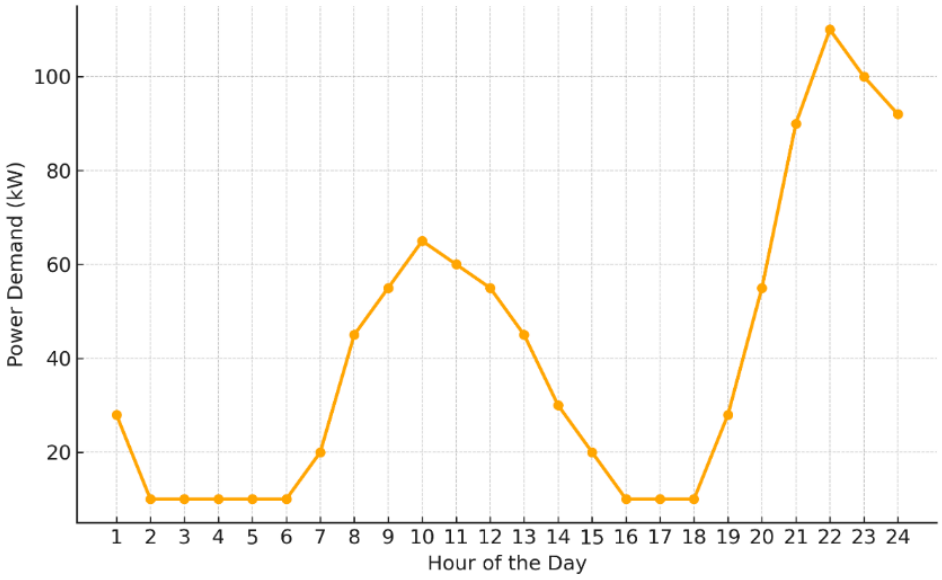



**Fig. 2.** Hourly profile of power demand in the studied microgrid.

### *3.4. PCC and Grid-Interaction Constraints*

Let $P_G(t)$ denote the net power exchange at the PCC (positive for **import**, negative for **export**). The interface capacity bounds the exchange,

$$-| P_G^{max} | \leq P_G(t) \leq | P_G^{max} | \quad (1)$$

and simultaneous import and export are forbidden. Equivalently, with $P_{imp}(t) \geq 0$, $P_{exp}(t) \geq 0$,

$$P_G(t) = P_{imp}(t) - P_{exp}(t) \quad (2)$$

$$P_{imp}(t)\, P_{exp}(t) = 0 \quad (3)$$

which enforces mutual exclusivity at each time step. In practice, anti-chattering is implemented via a small deadband around $P_G(t)=0$ and a dwell time before mode reversal; formal design and parameterization are provided with the control logic in Section 4 (CTDE multi-agent framework).

### *3.5. Network Topology and Power-Quality Considerations*

The linear LV feeder exhibits non-negligible impedance; under peak loading, end-node voltage drops can approach operational limits. Maintaining power quality therefore requires coordinated action from PV inverters, wind generation, diesel dispatch, and storage management. In all operating modes, the EMS targets a ±10 % voltage band at LV nodes, while respecting SOC envelopes, ramp-rate limits, and PCC import/export exclusivity. The detailed enforcement of these constraints (penalties and feasibility checks) is embedded in the control problem and will be specified alongside the objective/reward alignment (Section 4).

### *3.6. Sensing, Communication, and Data Layer*

An IoT-based sensing layer (smart meters, local controllers) continuously monitors generation outputs, nodal voltages, battery SOC, and PCC exchanges. These real-time measurements form the state vector consumed by the EMS. While the CTDE multi-agent design is introduced in Section 4, we note here that (i) independent local learners per resource (PV, wind, diesel, batteries) leverage locally available signals, and (ii) lightweight PCC coordination enforces exclusivity and anti-chattering without centralized actuation at execution time, consistent with the site's constraints and communication reliability [11]-[13], [24], [29], [37].

Based on this system description and operating constraints, the proposed control strategy is developed in the following section.

## 4. Control Strategy and Optimization Framework

This section presents the proposed CTDE multi-agent DQN framework, designed to ensure coordinated, grid-compliant and real-time operation.

Operating a hybrid microgrid in the Saharan context entails coping with volatile renewable resources, stochastic demand, and strict network-feasibility requirements on a long LV radial feeder. We adopt a decentralized, multi-objective control strategy under a centralized-training, decentralized-execution (CTDE) multi-agent design that targets economic, environmental, and power-quality objectives in real time [6], [11], [13], [17], [24], [27]-[29], [36], [37].

*4.1. Control Objectives and Composite Optimization*

The supervisory objective combines (i) monetary operating cost, (ii) $CO_2$ emissions from fossil units, (iii) reliance on the main grid at the PCC, and (iv) network-quality penalties (voltage/SOC). Over a horizon H (here H=24 steps), we minimize:

$$\min_{\{at\}} J = \sum_{t=1}^{H} [\underbrace{C_{imp}(t)\,[P_G(t)]^{+}\,\Delta t + C_{DG}\,P_{DG}(t)\,\Delta t}_{\text{monetary cost}} + \underbrace{\beta\, e_D\, P_D(t)\,\Delta t}_{CO_2} + \underbrace{\gamma\ |\,P_G(t)\,|\ \Delta t}_{\text{grid exchange}} + \underbrace{\delta\,\Phi(t)}_{\text{power quality}}\ ] \qquad (4)$$

Here $[x]^{+}=\max(x,0)$ (only imports are billed via $c_{mp}(t)$ ), and

$$\Phi(t) = \rho_V V_{viol}(t) + \rho_{SOC} SOC_{viol}(t) \qquad (5)$$

which aggregates per-step violations (in p.u.·h) of the LV voltage band (±10%) and the battery SOC envelope (see Sec. 3). The weights β, γ, δ are policy parameters explored in sensitivity studies (carbon, autonomy, power quality); by convention we set α=1 to anchor monetary units. This formulation is strictly aligned with the stepwise reward in Sec. 4.4 (the objective equals the negative cumulative reward) and is enforced under the physical and PCC constraints detailed in Sec. 3 and Sec. 4.3 (import/export exclusivity, deadband, dwell-time). See [4], [6]-[8] for treating the carbon term as a policy knob rather than a fixed national carbon price.

Conventions : $P_G(t)>0$: import; $P_G(t)<0$: export. The term $\gamma\,|P_G(t)|$ symmetrically penalizes reliance on the grid (autonomy), while $c_{imp}(t)\,[P_G(t)]^{+}$ captures the billing of imports. Δt is the time step (hours).

*4.2. CTDE Multi-Agent Architecture (Decentralized Execution)*

We adopt a Centralized Training with Decentralized Execution (CTDE) design with independent DQN learners per resource (PV, wind, diesel, batteries) and parameter sharing across homogeneous devices when applicable.

Training uses a centralized replay and joint state features (team context); execution is fully decentralized with per-agent greedy policies under action masking to enforce network feasibility.

*(a) Agents.*

- PV plants (3 agents). Each agent supervises one PV unit and selects discrete curtailment tiers and inverter modes, coordinating with local voltage/SOC/network constraints to avoid unnecessary exports during high-PV/low-load periods.
- Wind (1 agent). Selects discrete set-point/curtailment tiers consistent with feasible pitch/torque envelopes.
- Batteries (3 agents). Operate in {charge | hold | discharge} with discrete power tiers, subject to SOC bounds and converter ratings.
- Diesel (1 agent). Manages on/off commitment and discrete set-point tiers under ramp-rate limits and minimum up/down-time surrogates.
- PCC coordinator (1 agent, "policy-light"). Enforces import/export exclusivity, a zero-crossing deadband ζ, and dwell-times $T_{imp}^{min}$ , $T_{exp}^{min}$ (Eqs. (1)–(3), (18)–(19a and 19b)); selects tiered import/export levels within interface capacity and tariff logic.

*(b) State composition ($s_t$)*

Each agent observes (i) local signals (on-site generation estimate/forecast, local voltage, SOC, converter headroom), and (ii) broadcast context (aggregated demand, tariff, PCC mode, weather features). We use the following compact representation:

$s_t$=[ demand forecast slice, $P_{PV}(t)$, $P_{WT}(t)$, SOC(t), $P_G(t-1)$, $V_p(t)$, V summary, time-of-day, season, prices $c_{imp}(t)$ ].

with all features scaled; $V_{summary}$ denotes a reduced-order voltage descriptor (tail-bus proxy or PCA component).

*(c) Actions at*

Discrete, network-feasible choices per agent; invalid actions are masked (removed from the DQN argmax) to prevent violation-seeking behavior. Concretely:

- Diesel: {0,1} (off/on).
- Battery mode: {charge,hold,discharge} with clipped power.
- PCC: tiered import/export $\in\{-P_2,-P_1,0,+P_1,+P_2\}$ (levels within interface capacity, $P_G>0$ = import, $P_G<0$ = export).

*(d) Feasibility & masking.*

We mask actions that would violate: (i) exclusivity + deadband ($|P_G|<\zeta \Rightarrow P_G=0$), (ii) dwell-time (sign of $P_G$ cannot flip before $T_{dwell}$; implemented via $T_{imp}^{min}$ , $T_{exp}^{min}$, (iii) SOC bounds, device limits and ramp rates, and (iv) the LV voltage band (pre-check via a fast power-flow surrogate).

This keeps all executed actions network-feasible by construction and directly reflects field constraints [11], [13], [24].

*(e) Why discrete DQN here?*

Diesel on/off, battery modes (charge/hold/discharge) and tiered PCC exchange form a naturally discrete action space. DQN provides stable, sample-efficient, and edge-friendly learning under CTDE, with straightforward action masking for feasibility [23], [37].

Continuous-action DRL (DDPG/SAC/PPO) remains future work for set-point smoothing without altering the supervisory logic [28], [27], [38].

*(f) Note on PCC rules (consistency with Section 3.2b)*

The exclusivity/deadband/dwell logic used here matches the simulation evidence reported in Sec. 3.2(b): enforcing $\zeta$ and $T_{dwell}$ significantly reduces sign-flip attempts and micro-oscillations around zero exchange, stabilizing grid interaction while preserving import during deficits and export during surpluses.

This CTDE design mitigates the curse of dimensionality while retaining implementable local policies and a policy-tunable interaction with the main grid [27], [29], [36].

### *4.3. Embedded Physical and Network Models*

#### *4.3.1. Battery dynamics and health safeguards*

State-of-charge (SOC) evolves as:

$$SOC(t+1) = SOC(t) + \frac{\eta_c \cdot P_{ch}(t) - \frac{P_{dis}(t)}{\eta_d}}{E_{batt}} \cdot \Delta t \qquad (6)$$

Where:

- $Pch(t)$ and $Pdis(t)$ are the charging and discharging powers at time t (in W),
- $\eta c$ and $\eta d$ are the charging and discharging efficiencies, respectively,
- $Ebatt$ is the nominal energy capacity of the battery (in Wh),
- Δt is the time step duration (in hours).

Charging and discharging powers are bounded as follows:

$$0 \le P_{ch}(t) \le P_{ch}^{max} \qquad (7)$$

$$0 \le P_{dis}(t) \le P_{dis}^{max} \qquad (8)$$

To preserve lifetime and reserve, the SOC must remain within a secure operational range:

$$SOC_{min} \le SOC(t) \le SOC_{max} \qquad (9)$$

#### *4.3.2. LV feeder and voltage model*

In hybrid microgrids, ensuring high-quality service is not limited to supply–demand matching: voltage stability at all nodes must also be guaranteed. Excessive voltage deviations degrade sensitive equipment, reduce efficiency, and jeopardize reliability, issues amplified in Saharan feeders where distances are long and renewable injections are intermittent [12], [13].

We adopt a realistic nodal topology consistent with prior work on downstream DER integration in distribution networks [12], [13]. Node voltages are computed by aggregating voltage drops along feeders as functions of active/reactive flows, segment lengths, and line parameters. Compliance with national practice is enforced by constraining voltages within ±10% of nominal across all operating conditions [12], [13].

Fig. 3. illustrates the simplified single-line diagram used to index consumers and Distributed Energy Resource (DERs). Let n1 denote the number of consumers upstream of $DER_1$; $n_j$ the consumers between $DER_{j-1}$ and $DER_j$; $n_{M-1}$ those between $DER_{M-2}$ and $DER_{M-1}$; and $n_{M_n}$ the consumers between $DER_{M-1}$ and the feeder tail (a fictitious $DER_M$ with zero output is assumed if none is installed). With N the total consumers and M the number of DERs, we have

$$N = \sum_{j=1}^{M} n_j \qquad (10)$$

For a line segment of length L (m), linear resistance r (Ω/m) and reactance x (Ω/m), supplying active power P (W) and reactive power Q (VAR) at voltage U (V), the voltage drop is:

$$\Delta U = \frac{L}{U}(rP + xQ) \qquad (11)$$

with nodal compliance:

$$U_{min} \le U_{j,i}(t) \le U_{max} \qquad (12)$$

with $\frac{U_{min}}{U_{max}} = (1 \mp 0.10)U_s$

Where, $U_s$ denotes the contractual (service) nominal voltage at the relevant supply level

These relations are embedded in the environment and in the penalty structure so that learned policies remain network-feasible [12], [13], [24], [37].

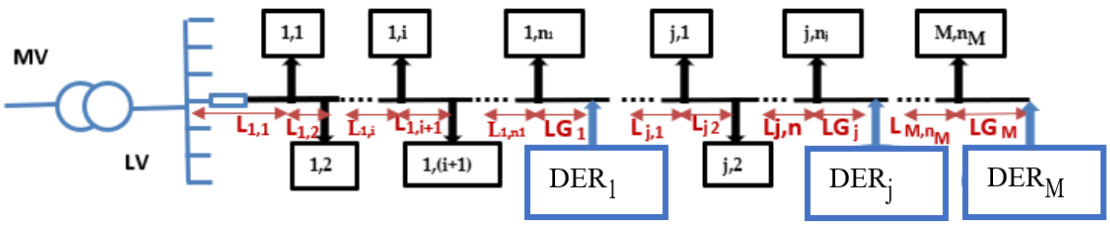



**Fig. 3.** Simplified line diagram showing network nodes, consumer connection points, and distributed generation units [13].

### *4.4. Economic Cost, Emissions and Instantaneous Reward*

*a) Monetary cost (per step, Δt = 1 h)*

$$C_{total}(t) = \Delta t(|P_B(t)| \cdot C_B(t) + P_{DG}(t) \cdot C_{DG}(t) + [PG(t)]^{+} \cdot C_{imp}(t)) + C_{sec}(t) \qquad (13)$$

Where, $[x]^{+}=\max(x,0)$ charges import only; $P_B(t)$ is the signed battery power (discharge/charge), $P_G(t)$ the PCC exchange (positive = import), $P_{DG}(t)$ the diesel generator output, $C_{imp}(t)$ the grid tariff, and $C_{sec}(t)$ aggregates penalty costs due to curtailment, unserved energy, and power-quality

violations (units: MAD per time step Δt). $C_{DG}(t)$ denotes the diesel generator marginal cost (fuel + variable O&M) and $C_B(t)$ the battery cycling cost per unit throughput.

In our setting, grid export is not remunerated; therefore, only imports are billed via $[PG(t)]^+ C_{imp}(t)$. If an export tariff applies, include a revenue term $-P_{exp}(t)C_{exp}(t)$.

*b) Emissions (per step)*

$$E_{CO_2}(t) = e_{DG}\, P_{DG}(t)\, \Delta t \quad (14)$$

With, $e_{DG}$=0.8 kg/kWh as the diesel emission factor (consistent with the 24-hour energy and 272.5 kg $CO_2$ balance reported later).

*c) Reward shaping (instantaneous)*

$$R_t = -(\alpha\, C_{total}(t) + \beta\, E_{CO_2}(t) + \gamma\, [P_G(t)]^+ \Delta t) - \delta\, \Phi(t) \quad (15)$$

$$\Phi(t) = \rho_V V_{viol}(t) + \rho_{SOC} SOC_{viol}(t) + \rho_{curt} Curt(t) + \rho_{uns} Unserved(t) \quad (16)$$

The planning objective over horizon T is :

$$J = \sum_{t=1}^{T} (\alpha C_{total}(t) + \beta E_{CO2}(t) + \gamma([PG(t)]^+).\Delta t + \delta\Phi(t)) \quad (17)$$

Maximizing the cumulative return $\sum_{t=1}^{T} R_t$ is equivalent to minimizing the objective J. Training optimizes the normalized, unitless reward defined in Eqs. (22)–(28). The economic objective in Eq. (4) is for reporting and interpretation only and is numerically equal to the negative cumulative return; hence, Eqs. (15)–(17) are **not** used during training to avoid unit inconsistencies. In short, Eqs. (15)–(17) provide a raw, dimensional multi-objective formulation for reporting**,** whereas learning is driven by the scale-normalized reward in Eqs. (22)–(28), which prevents unit coupling, stabilizes training, and clarifies trade-offs across scenarios.

*d) Time step convention.*

Unless otherwise stated, all per-step quantities use Δt=1h. PCC-specific toggling metrics in Table 1 are computed at 15-min granularity for finer chattering diagnosis.

### *4.5. Grid Interaction: PCC Coordination and Feasibility*

Let $P_G(t)$ denote net PCC exchange (**>0** import, **<0** export). Capacity and mutual exclusivity:

- $-|P_G^{max}| \le P_G(t) \le |P_G^{max}|$ ;
- $P_G(t) = P_{imp}(t) - P_{exp}(t)$ ;
- $P_{imp}(t)\, P_{exp}(t) = 0;\quad P_{imp}(t) \ge 0;\ P_{exp}(t) \ge 0.$

*(a) Anti-chattering deadband and dwell times:*

$$|\, PG(t)\,| \in \{0\} \cup [\zeta, P_G^{max}] \quad (18)$$

with a small deadband $\zeta$>0 around zero exchange to prevent rapid import/export toggling.

and binary mode flags $z_{imp}(t), z_{exp}(t) \in \{0,1\}$ with $z_{imp}(t) + z_{exp}(t) \le 1$ minimum up/down times $T_{imp}^{min}, T_{exp}^{min}$ enforce dwell-time (UC-style) to limit switching wear and oscillations:

once a mode is entered, it must persist for at least $T_{mode}^{min}$

*(b) Reactive exchange / power factor:*

$$PF(t) = \frac{|P_G(t)|}{\sqrt{P_G(t)^2 + Q_G(t)^2}} \ge PF_{min} \quad (19a)$$

With ,

$$|Q_G(t)| \le \tan(\arccos(PF_{min})).|P_G(t)| \quad (19b)$$

These rules are implemented by the PCC coordinator and shared as state flags, ensuring grid-compliant exchange [11]-[13], [24]. Their stabilizing effect (reduced sign flips, fewer micro-oscillations near zero) matches the simulation evidence reported in Sec. 3.2(b).

*(c) Power Balance and Operational Constraints*

$$P_{PV}(t) + P_{WT}(t) + P_{DG}(t) + P_G(t) + P_B(t) = P_{load}(t) + Curt(t) + P_{loss}(t) \quad (20)$$

with device nameplate/rating limits, diesel ramp-rates, and battery constraints (Eqs. (5)–(8), (20)). Any violation triggers the corresponding terms in Φ(t), steering the policy toward secure dispatch.

Table 1 quantifies the impact of the deadband ζ and dwell time τ on PCC behavior over a 24 h episode (Δt = 15 min).

**Table 1.** PCC exclusivity metrics with and without enforcement of deadband and dwell-time (Δt = 15 min)

| Metric | Tiered only | Enforced (deadband + dwell) |
|---|---|---|
| Sign toggles (±) | 2 | 2 |
| Blocked sign-change steps (dwell) | 0 | 13 |
| Import energy (kWh) | 375 | 350 |
| Export energy (kWh) | 243.8 | 181.2 |
| Zero steps (count) | 18 | 31 |

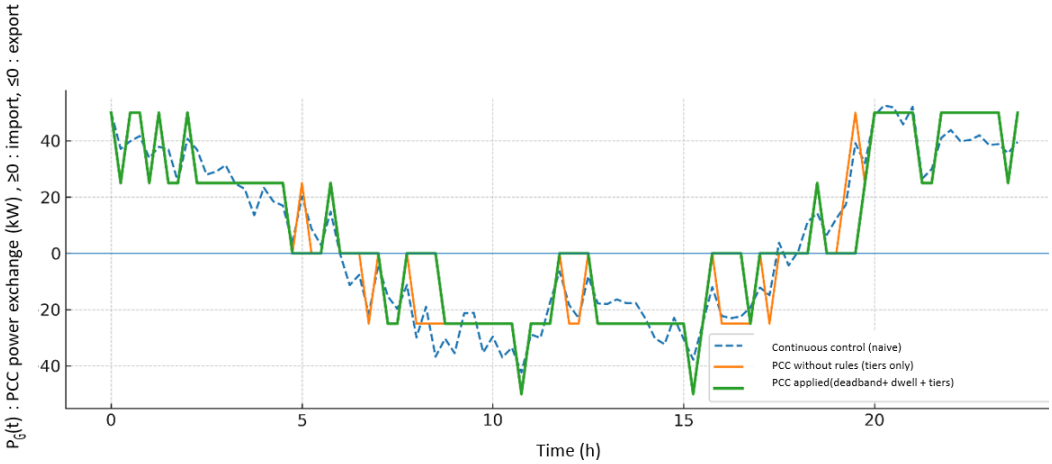



**Fig. 4.** Effect of the deadband ζ around zero PCC exchange.

Fig. 4. illustrates how the deadband ζ around zero and the dwell time τ suppress sign flips and enforce import/export exclusivity at the PCC.

### *4.6. Stochastic Environment, Episodes and Training Protocol*

We simulate hourly steps over 24-h episodes, under summer/winter weather regimes (irradiance, wind) and stochastic demand with realistic diurnal structure. The agent uses experience replay and a target network; exploration follows decaying ε-greedy; the Q-network is an MLP with ReLU activations, optimized with Adam [13], [23], [25], [27], [37], [38].

## 5. Methodology Based on Deep Q-Network (DQN)

### *5.1. DQN Overview*

The supervisory controller builds on the Deep Q-Network (DQN) paradigm, which couples value-based reinforcement learning with deep function approximation to learn near-optimal dispatch policies in high-dimensional, partially uncertain environments without requiring explicit analytic models of full system dynamics [17], [20], [28]. A neural approximator $Q_\theta(s,a)$ represents the action-value function, where the state s encodes the instantaneous microgrid condition and the action a maps to a physically admissible control decision. The optimality equation is

$$Q^*(s,a) = E[r + \gamma_{RL} \underbrace{\max_{a'}} Q^*(s',a')] \quad (21)$$

with discount factor $\gamma_{RL} \in [0,1]$ trading off immediate and future returns. In practice, DQN updates θ by minimizing the squared temporal-difference error between $Q_\theta(s,a)$ and a target $y = r + \gamma_{RL} \max_{a'} Q^*_{\bar{\theta}}(s',a')$, where $\bar{\theta}$ denotes a slowly updated target network, a stabilization device widely adopted in power-system DRL [21], [24], [25]. Together with experience replay (random mini-batches from a transition buffer), these mechanisms decorrelate updates and mitigate divergence, enabling generalization under the non-stationary regimes typical of renewable-dominated microgrids [6], [13], [27], [28], [29].

### *5.2. CTDE Multi-agent Architecture, Data Flow, and Environment Loop*

The proposed control framework follows a Centralized Training, Decentralized Execution (CTDE) paradigm, combining independent deep reinforcement-learning agents with a lightweight Point of Common Coupling (PCC) coordinator.

During training, agents exploit global information - including electricity tariffs, grid availability, nodal voltages, and short-term forecasts - whereas during execution, decisions rely strictly on local observations, in accordance with realistic field deployment constraints.

The PCC coordinator enforces import/export exclusivity, zero-crossing deadbands, and minimum dwell times, ensuring grid-compliant operation without compromising decentralized autonomy.

Figs. 5-7 collectively illustrate the CTDE architecture and its data flow.

Fig. 5 depicts the centralized training pipeline, where global signals are aggregated to construct the joint state. Exploration is handled via an ε-greedy policy, transitions are stored in a replay buffer, and policy updates are performed using minibatch stochastic gradient descent (SGD) with target networks for stabilization.

Fig. 6 details the CTDE coupling mechanism, emphasizing the separation between global information available during training and the local-only observations accessible at execution time. Feasibility checks and action masking are systematically applied to prevent violations of device and network constraints.

Fig. 7 summarizes the decentralized execution layer, in which independent local agents - associated with PV, wind, diesel, and battery resources - select discrete control actions. A policy-light PCC coordinator enforces interface rules and exposes operational mode flags, while device-level enforcement enables low-latency edge inference [11]-[13], [24], [27], [37].

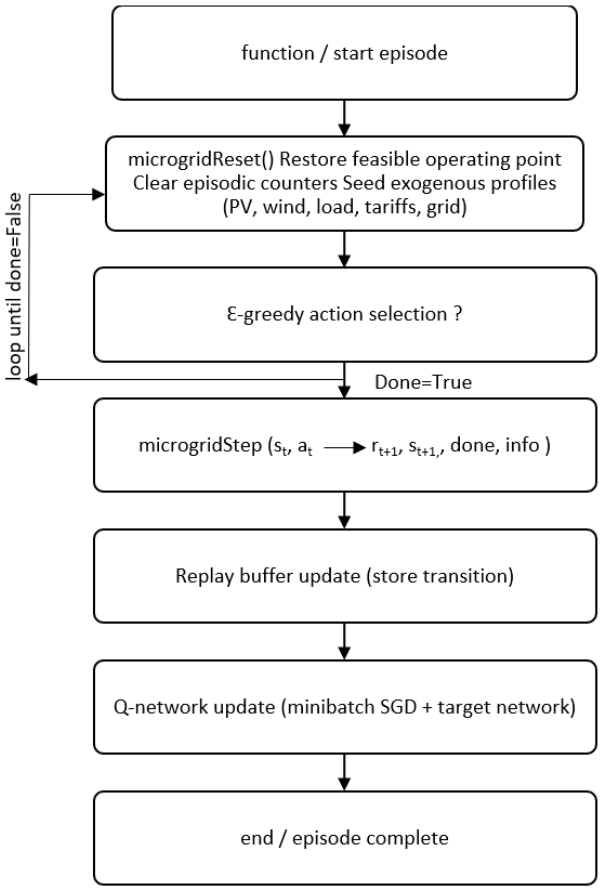



**Fig. 5.** Centralized training pipeline with global signals (tariffs, grid, voltages, forecasts); replay/target networks and minibatch SGD

The environment dynamics are governed by the microgrid-Step routine illustrated in Fig. 8. At each time step, the current state st is read, the selected joint action at is applied, and device-, battery-, and network-level limits are enforced through feasibility filtering and action masking. A scalar reward rt+1 is then computed, and the system transitions to the next state st+1, with terminal flags indicating episode completion when applicable.

Episode initialization and lifecycle management are handled by the microgrid-Reset procedure shown in Fig. 9. This routine restores a feasible operating point, clears episodic counters, and seeds all exogenous profiles, including PV generation, wind availability, load demand, dynamic tariffs, and grid connection status. The resulting episodic loop underpins both centralized training and decentralized execution.

Regarding agent roles, three PV agents actuate discrete inverter and curtailment tiers, while the wind agent selects feasible set-point or curtailment levels. Three battery agents operate over the discrete action set {charge | hold | discharge}, subject to state-of-charge and converter constraints, with tiered power levels. The diesel agent manages on/off commitment and discrete set-point tiers, incorporating ramping limits and minimum up/down-time surrogates. Throughout operation, the PCC coordinator applies exclusivity rules, deadband constraints, and dwell-time enforcement at the interface level via environment-level constraints and action masking. This design preserves decentralized decision-making while ensuring stable, grid-compliant power exchanges [11]-[13], [24], [27], [37].

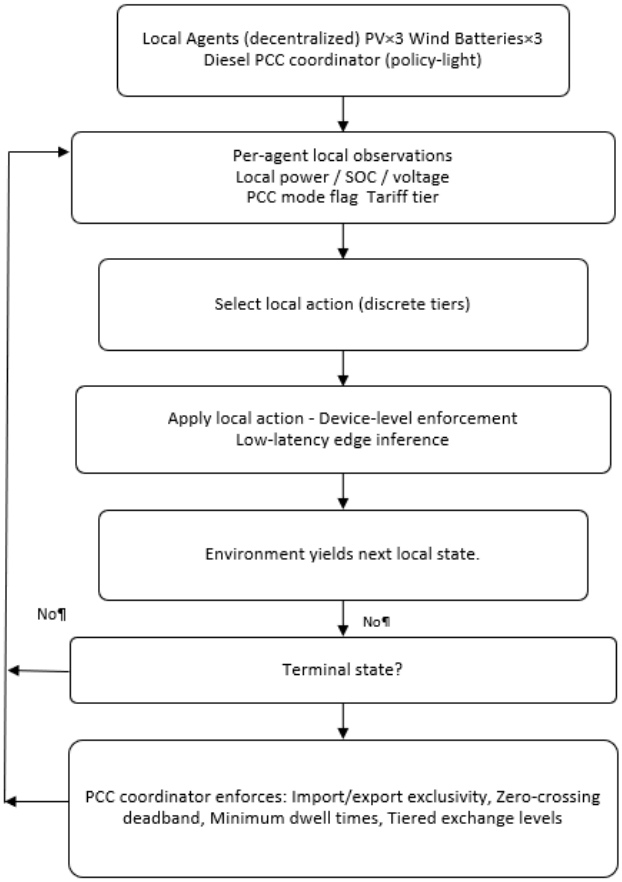



**Fig. 6**. CTDE coupling: global information for training vs. local-only observations at execution; feasibility checks and action masking.

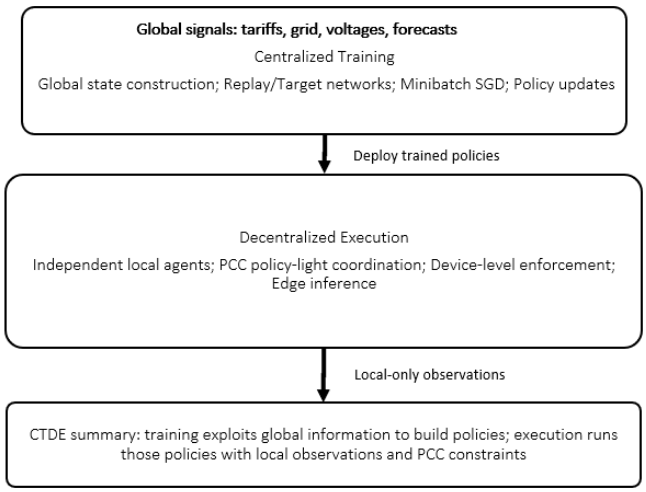



**Fig. 7.** Decentralized execution: independent local agents plus PCC coordination; device-level enforcement; edge inference.

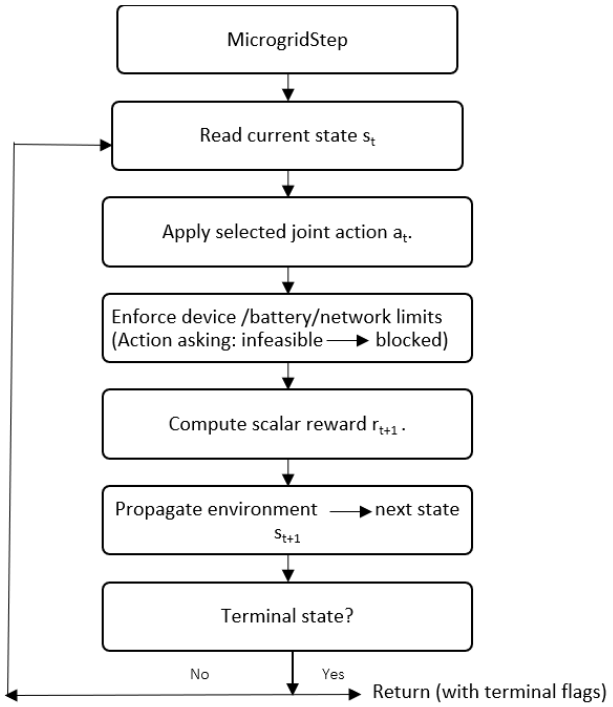



**Fig. 8.** microgrid-Step: read state → apply joint action → enforce limits (mask if infeasible) → compute reward → propagate to next state with terminal flags.

### 5.3. *State, Action, and Reward Design*

We adopt a CTDE multi-agent design with independent DQN learners per resource. The interface is defined as follows:

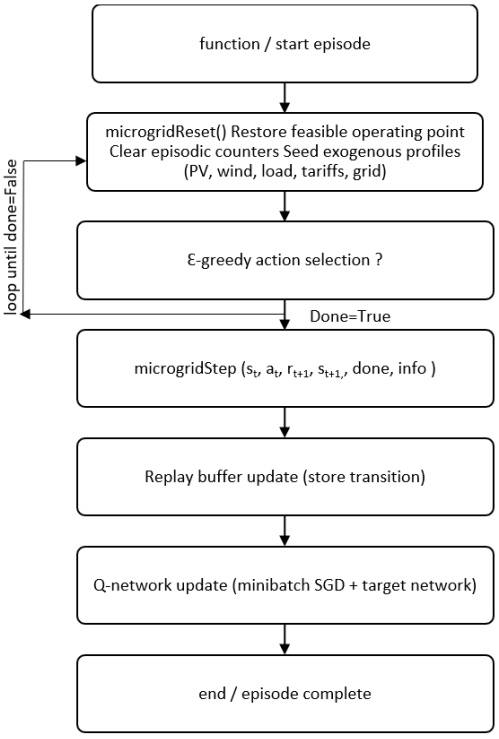



**Fig. 9.** Microgrid-Reset and episodic loop: restore feasible operating point, clear counters, seed exogenous profiles.

*a) State vector*

Joint state aggregates measured/estimated variables: PV and wind outputs (and forecasts when available), aggregate and nodal loads, battery SOC, node-wise voltages along the 31-node feeder, PCC import/export status and tariff tier, and device headrooms. Each agent receives its local slice plus a compact broadcast context (e.g., tariff tier and PCC mode).

*b) Action space*

Discrete, hardware-consistent actuation: diesel on/off; battery mode tiers (charge/hold/discharge within ratings); PV/wind curtailment tiers; tiered grid exchange at the PCC. This discreteness matches contactor-level logic and tariff steps and motivates a DQN family over continuous-action methods; continuous DRL (DDPG/SAC/PPO) is reserved for future fine set-point tuning. [21], [24], [26], [27].

*c) Reward shaping*

$$R_t = -\left(\alpha\,\tilde{C}(t) + \beta\,\tilde{E}_{CO_2}(t) + \gamma\,\tilde{P}_G(t) + \delta\,\tilde{V}_{viol}(t)\right) \quad (22)$$

With

$$\tilde{C}(t) = \frac{C_{total}(t)}{C_{ref}};\ \tilde{E}_{CO_2}(t) = \frac{E_{CO_2}(t)}{E_{ref}};\ \tilde{P}_G(t) = \frac{|P_G(t)|}{P_{ref}};\ \tilde{V}_{viol}(t) = \lambda_V \Psi_V(t) + \lambda_{soc} \Psi_{soc}(t) \quad (23)$$

where $C_{ref}$, $E_{ref}$, Pref normalize heterogeneous units (MAD, $tCO_2$, kW) for learning stability, and $\tilde{V}_{viol}(t)$ aggregates voltage and SOC penalties using internal weights $\lambda_V$, $\lambda_{soc}$ (not policy weights).

*d) Voltage and SOC penalties follow*

$$\Psi_V(t) = \sum_i [\max(0, \frac{|V_i(t) - V_{nom}|}{V_{nom}} - \Delta_{tol})]^2 \quad (24)$$

$$\Psi_{soc}(t) = [\max(0,\ SOC_{min} - SOC(t)) + \max(0,\ SOC(t) - SOC_{max})]^2 \quad (25)$$

With, $\Delta_{tol}$=0.10 p.u. (±10%).

***e)*** *Degradation & autonomy notes*

Battery degradation is not double-counted: it is monetized via $C_B(t)$ inside $C_{total}(t)$ (Section 4, Eq. (13)) rather than a separate reward term. Likewise, an extra autonomy penalty $\Phi_{pcc}(t)$ is omitted by default because import is already captured by the $\gamma\,\tilde{P}_{import}(t)$ term; it can be activated in dedicated autonomy-first experiments [8], [11], [14], [24], [37].

### 5.4. *Training Protocol and Hyperparameters*

A physics-constrained digital twin calibrated to Boujdour reproduces stochastic PV/wind resources, the hourly demand structure, battery round-trip efficiencies and SOC envelopes, and node-wise voltages along the feeder [10], [11], [14], [15]. Training uses two representative 24-h episodes (summer / winter), each with 24 steps, repeated 500–1000 times under a decaying ε-greedy exploration (1.0 → 0.01). Experience is stored in a replay buffer of 10 000 transitions; i.i.d (independent and identically distributed). minibatches of 64 are sampled at each update. The Q-approximator is a 2-layer MLP (64/32, ReLU) optimized with Adam (learning rate $5\times10^{-4}$). Stabilization relies on a target network (periodic hard

syncs or equivalent soft updates) and the replay mechanism; the discount is $\gamma_{RL}$ = 0.99 [6], [13], [21], [24], [25], [27]. Unless stated otherwise, per-step quantities use $\Delta t$ = 1 h; both seasonal scenarios are shuffled within the training loop. (Full per-agent action spaces are identical except for device-specific tiers; the complete configuration can be found in Supplementary S3.)

### 5.5. *Reward Formulation and Normalization*

At each time step t, we compute three raw metrics: the monetary cost C(t) (MAD), the instantaneous $CO_2$ emissions E(t) (kg), and a power-quality/constraint metric Φ(t) (e.g., p.u.·h for voltage violations or a composite score). To avoid unit inconsistencies and to stabilize learning, we train the agent with a unitless reward obtained by normalizing each raw metric by a positive calibration scale:

$$\tilde{C}(t) = \frac{C(t)}{S_C}\ ;\ \tilde{E}(t) = \frac{E(t)}{S_E}\quad ;\tilde{\Phi}(t) = \frac{\Phi(t)}{S_\Phi}$$

where $S_C$ (MAD), $S_E$ (kg), and $S_\Phi$ (p.u.·h or dimensionless) are fixed scales defined below.

*a) The training reward is then*

$$R_t = -(\beta_1 \tilde{C}(t) + \gamma_1 \tilde{E}(t) + \delta_1 \tilde{\Phi}(t)) \qquad (26)$$

with $\beta_1$, $\gamma_1$, $\delta_1 \geq 0$ dimensionless trade-off weights. The agent maximizes the discounted return

$$G_t = \sum_{k=0}^{T-t-1} \rho^k R_{t+k+1} \quad ;\ 0 < \rho \leq 1 \qquad (27)$$

where ρ is the discount factor (we deliberately avoid using γ here to prevent confusion with the emission weight γ above).

*b) Calibration of normalization scales.*

We adopt a pragmatic calibration to make unitless quantities interpretable:

- *Option A (baseline-relative):*

$S_C = C_{baseline}$, $S_E = E_{baseline}$, $S_\Phi = \Phi_{baseline}$ , where "baseline" denotes a non-RL policy (e.g., heuristic or PSO). This makes $\tilde{\cdot} \approx 1$ represent baseline performance.

- *Option B (bounded by physics/limits):*

$S_C$, $S_E$, $S_\Phi$ chosen as credible per-step upper bounds (e.g., tariff-at-peak, diesel at rated power, maximum allowed violation). This typically yields $\tilde{\cdot} \in [0,1]$.

In our experiments, we use option A and report the actual values of $S_C$, $S_E$, $S_\Phi$ in the Appendix/Experimental Setup for reproducibility.

*c) Constraints handling.*

Physical and network-feasibility constraints - including PCC power limits, battery state-of-charge (SOC) bounds, voltage and power-factor requirements, as well as dwell-time and deadband constraints for power sign changes - are enforced through a dual mechanism: (i) action masking to prevent infeasible decisions whenever applicable, and (ii) a bounded penalty term incorporated into the reward function Φ(t) to account for any residual constraint violation. A generic composite formulation is given by

$$\Phi(t) = w_V \Phi_V(t) + w_{PF} \Phi_{PF}(t) + w_{PCC} \Phi_{PCC}(t) \qquad (28)$$

where each penalty component $\Phi_i(t) \geq 0$ is equal to zero when the corresponding constraint is satisfied and increases monotonically with the magnitude and/or duration of the violation. The weighting coefficients $w_i$ regulate the relative contribution of each constraint term and are kept constant across all policies and simulation scenarios to ensure fair comparison.

*d) Reporting vs. raining*

Training minimizes the unitless aggregate $\beta\tilde{C} + \gamma\tilde{E} + \delta\tilde{\Phi}$ through reward maximization. For interpretability, all results are reported in raw KPIs: total cost $\sum_t C(t)$(MAD), total emissions $\sum_t E(t)$ (kg), grid import/export (kWh), and voltage violation hours $V_{viol}$ (p.u.·h). The descriptive *economic* objective -(for reporting only; training uses the normalized reward)- shown elsewhere mirrors the reward structure but is not used directly during learning.

### 5.6. *Implementation and Co-simulation*

The DQN agents are implemented in Python (TensorFlow) and co-simulated with a detailed MATLAB microgrid model to ensure consistent feeder power-flow responses and device-level constraints. This modular design clarifies constraint handling (voltage and SOC), supports edge deployment for low-latency inference, and follows recent grid-edge DRL practices [6], [24], [26], [27], [37]. Batteries are monitored with SOC-centric safeguards aligned with contemporary estimation/management practices to preserve lifetime under frequent cycling [14], [15].

## 6. Results and Discussion

The performance of the proposed framework is evaluated under realistic operating scenarios, as detailed in the following subsections.

Although the agent is trained with a unitless normalized reward, all performance is reported using raw KPIs for interpretability: total operating cost (MAD), $CO_2$ emissions (kg), grid import/export (kWh), and voltage-violation hours (p.u.·h). Unless stated otherwise, improvements are given as absolute values and relative percentages with respect to the baseline.

### 6.1. *Evaluation Setup*

We evaluate the CTDE multi-agent DQN on a 24-h digital-twin of the Boujdour microgrid (summer/winter profiles), with stochastic PV/wind and a diurnal load consistent with §3. The environment enforces device/network constraints (SOC, ramp-rates, PCC exclusivity & dwell, ±10% voltage band) and computes the policy reward aligned with the multi-objective formulation in Eq. (4) (cost–$CO_2$–import–quality) [2], [5]-[8], [16], [18]-[20], [24], [27], [29], [37].

### 6.2. *Baseline Multi-objective Performance (24-h)*

Energy mix & autonomy. Over a 24-h horizon with total demand 991.1 kWh, the learned dispatch is renewables-first: PV 559.4 kWh (56.4 %), diesel 340.6 kWh (34.4 %), and main-grid exchange 91.2 kWh (9.2 %). The batteries act as

shallow buffers with a throughput of ± 7.6 kWh, indicating that most balancing is achieved via coordinated PV curtailment tiers, diesel commitment tiers, and tiered PCC exchange, rather than deep cycling.

Emissions. With a diesel factor of 0.8 kg kWh$^{-1}$, total $CO_2$ amounts to 272.5 kg, i.e., an intensity of ≈ 275 g kWh$^{-1}$ across the day: consistent with the reductions typically reported for learning-based EMS in hybrid systems [4], [6].

Takeaway. PV is prioritized; diesel covers edge hours (early morning/evening); the grid remains tertiary. This mix delivers limited grid reliance and restrained cycling while maintaining feasibility and service quality. (A full numeric breakdown formerly shown in Table 4 is retained in Supplementary S4.)

### *6.3. Temporal Dispatch and Interpretability*

Overview. A single multi-panel view (Figs. 10–13) summarizes source-wise dispatch, balance with demand, carbon effects, and battery behavior over 24 h.

Source-wise dispatch (fig. 10). The three PV arrays dominate at mid-day; wind ramps in late afternoon; the diesel unit covers early/late deficits with short, targeted runs.

Generation vs. load (Fig. 11). Two demand peaks (≈ 11:00; 20:00–22:00) are aligned with selective diesel commitment and shallow battery support, limiting over-production and import spikes.

Emissions profile (Fig. 12). Hourly $CO_2$ remains low during PV-rich periods; controlled battery discharge mitigates evening emissions by shaving peaks.

SOC dynamics (Fig. 13). All three units stay within [$SOC_{min}$,$SOC_{max}$], with heterogeneous roles: one daily cycler, one peak-shaver, and one buffer, a merit-order pattern typical of value-based DRL policies [24], [37].

The panels collectively show a renewables-first dispatch with restrained cycling, grid stability at the PCC, and interpretable battery roles that generalize across operating conditions.

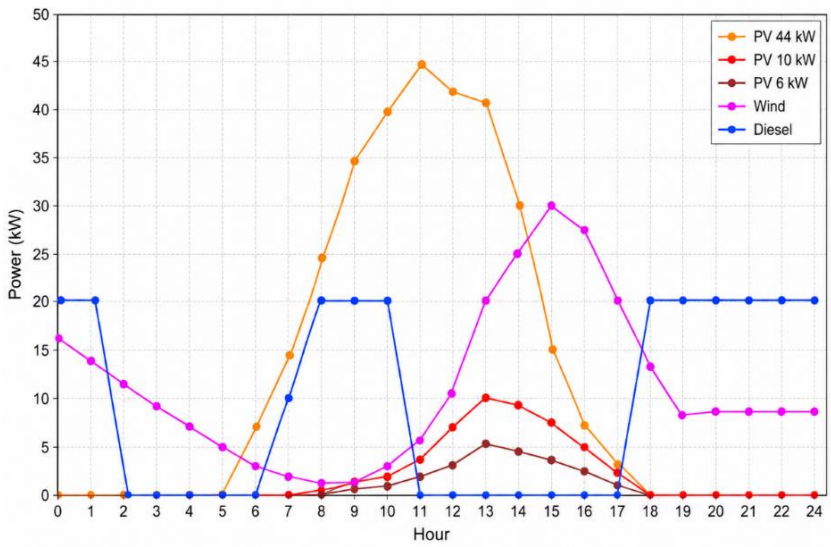



**Fig. 10.** Source-wise power (PV-44 kW, PV-10 kW, PV-6 kW, wind, diesel).

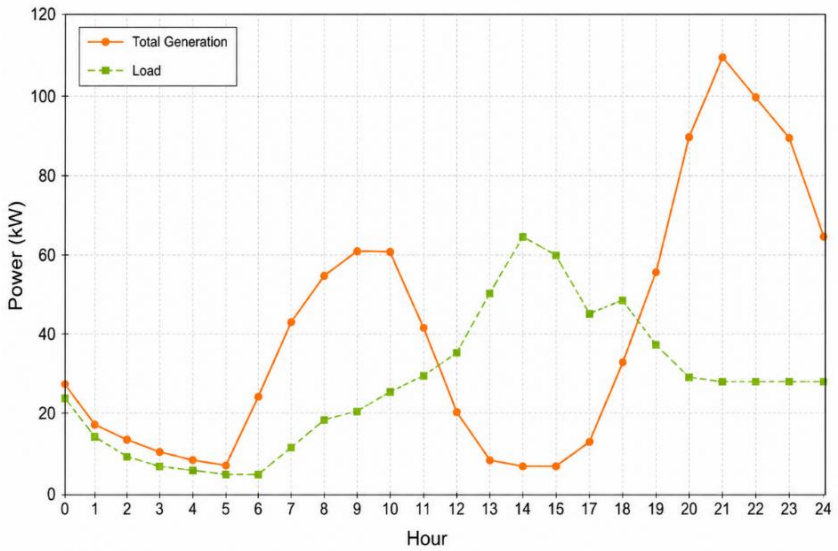



**Fig. 11.** Total generation vs. load.

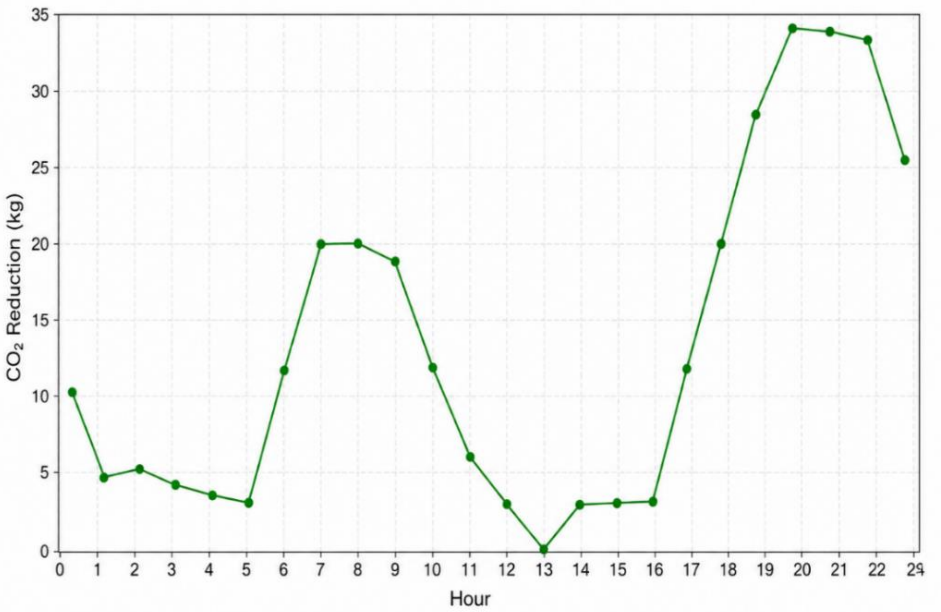



**Fig. 12.** Hourly $CO_2$ reduction attributable to the learned policy (relative to diesel-centric operation)

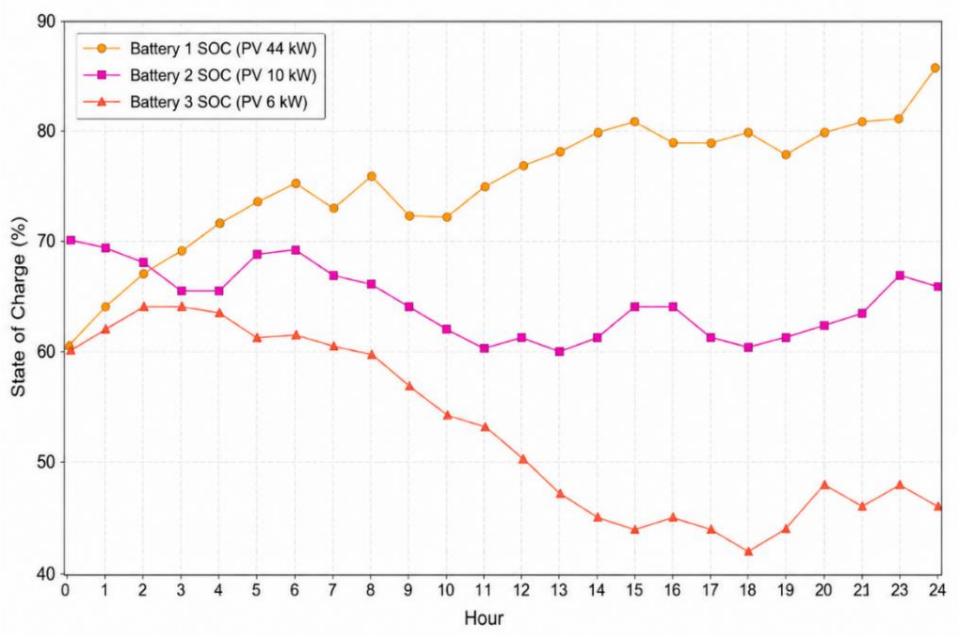



**Fig. 13.** Battery SOC trajectories (three PV-coupled units), all within operational bounds; roles emerge as daily cycler, peak-shaver, buffer [24], [37].

### *6.4. Grid-interaction Behavior (PCC)*

The PCC coordinator enforces import/export exclusivity, a zero-crossing deadband and dwell times, which together suppress near-zero oscillations and prevent chatter. In practice we observe few reversals and stable tiered exchanges aligned with tariffs and local balance, an expected result of coupling cost/import weights with voltage-penalty terms in the reward and feasibility masks [11]-[13], [24]. Quantitative effects (blocked sign-change attempts, increase in zero-exchange steps, reduced gross import/export) are reported in Sec. 3.2(b) and visualized in Fig. 4.

### *6.5. Voltage-quality Compliance on the LV Feeder*

Voltage profiles along the low-voltage (LV) feeder remain fully compliant with the ±10% admissible band, despite the presence of long radial sections and intermittent power injections from distributed energy resources [12], [13]. The time-series voltage trajectories observed at representative buses - namely nodes 1, 15, and 31 - exhibit smooth and stable behavior throughout the 24-hour horizon, with no noticeable excursions beyond the allowable limits, as illustrated in Fig. 14.

This behavior is further confirmed by the spatio-temporal voltage surface spanning all monitored nodes and hourly time

steps. As shown in Fig. 15, voltage variations evolve gradually across both space and time, without abrupt drops or deep localized sags, indicating a well-regulated feeder under the proposed control strategy.

Consistently, the aggregated voltage-violation metric $V_{viol}(t)$ remains close to zero over the entire horizon. This outcome reflects the effectiveness of the enforced power-factor constraint and the voltage-aware penalty structure embedded in the learning process, which together internalize network feasibility and ensure stable LV operation during both training and execution phases [12], [13], [24], [37].

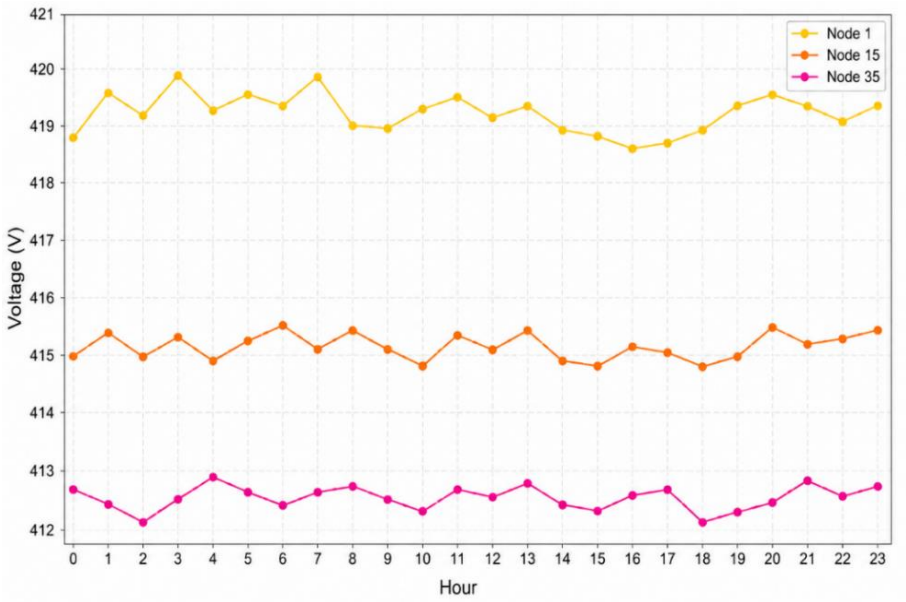



**Fig. 14.** LV voltage quality assessment.

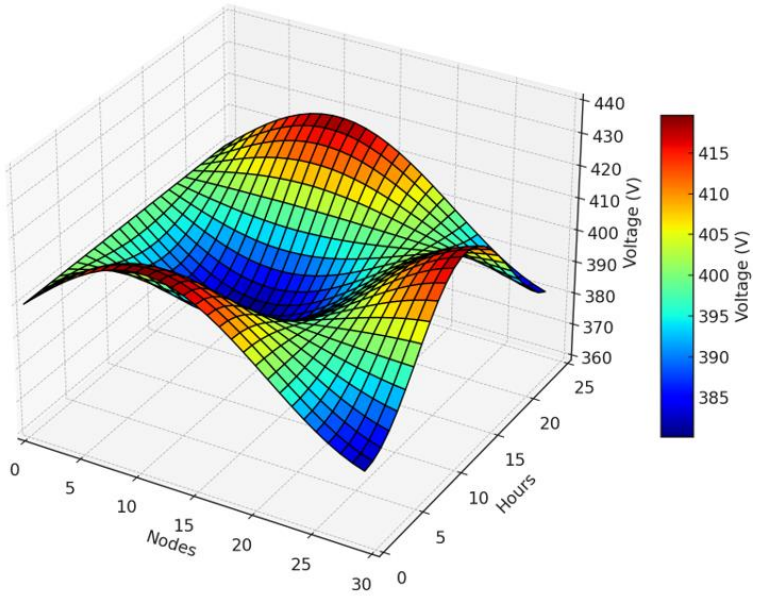



**Fig. 15**. LV voltage quality assessment.

Voltage at selected nodes (1, 15, and 31) over 24 h, remaining within the ±10% compliance band.

Spatio-temporal voltage surface (nodes × hours), showing smooth voltage variations and the absence of deep sags.

### 6.6. *Sensitivity to Policy Weights*

#### 6.6.1. *$CO_2$ penalty weight β*

We sweep β∈{0.1,0.5,1.0,2.0} with the other weights fixed. The trend is monotonic: a larger β reduces diesel use and $CO_2$ at the expense of higher cost. Concretely, for the same 24-h load (991.1 kWh):

- Cost (MAD): 1100 → 1135 → 1180 → 1270 (+15.5%)
- $CO_2$ (kg): 310 → 290 → 265 → 240 (−22.6%), i.e., ≈313 → ≈242 g $kWh^{-1}$

- Diesel (kWh): 360 → 340 → 320 → 290 (−19.4%)

This behavior reflects the reward trade-off encoded by β and aligns with policy-tuning effects reported for value-based DRL in hybrid systems [24], [37].

*6.6.2. Import reliance γ and quality weight δ*

The sensitivity of the proposed control framework to the import-reliance weight γ and the power-quality weight δ is investigated through systematic parametric sweeps. These analyses aim to clarify how grid dependency, emissions, voltage quality, and operational cost trade off against one another under multi-objective reinforcement learning.

Across the ($\beta$,$\gamma$) sweep, increasing the import-reliance weight γ progressively penalizes grid exchanges, resulting in a marked reduction of the absolute grid power $|P_G|$. This shift forces a greater reliance on local balancing resources - namely diesel generation and battery storage - which slightly increases the total operating cost. However, within a moderate γ range, this cost increase remains limited while grid dependency is significantly reduced.

Similarly, across the (β,δ) sweep, increasing the power-quality weight δ enforces stricter voltage-compliance constraints, reflected by lower accumulated voltage-violation severity $V_{viol}$. This improvement is typically achieved through additional curtailment or more conservative dispatch decisions, again with only modest cost penalties when δ is tuned within a practical range [24], [37].

A coherent and practically relevant operating region emerges around β ≈ 0.5-1.0 and γ ≈ 0.1-0.2, where total cost remains moderate while both grid imports and $CO_2$ emissions are effectively reduced. This region represents a balanced compromise between economic performance, autonomy, and environmental impact.

Table 2 summarizes three Pareto-efficient operating points selected from the combined (β, γ, δ) sweeps, each emphasizing a different operational priority: carbon reduction, grid autonomy, or voltage quality. For each point, the table reports the total cost (MAD), $CO_2$ emissions (kg), absolute grid exchange $|P_G|$. (kWh), and voltage-violation severity $V_{viol}$.(p.u.h.).

**Table 2**. Pareto-efficient operating points **(β,γ,δ).**

| **(β,γ,δ)** | **Cost (MAD)** | **$CO_2$ (kg)** | **$\|P_G\|$ (kWh)** | **$V_{viol}$ (p.u.·h)** | **Notes** |
|---|---|---|---|---|---|
| (2.0, 0.10, 100) | 1278 | 241 | 122 | 0.28 | Carbon-lean (minimum emissions at moderate grid use) |
| (0.5, 0.30, 100) | 1186 | 289 | 102 | 0.34 | Autonomy-lean (low grid exchange with moderate cost) |
| (0.5, 0.10, 400) | 1142 | 292 | 128 | 0.21 | Power-quality-lean (tight voltage via higher δ) |

Figs.16-18 present two-dimensional sweeps over (β,γ) for total operating cost, $CO_2$ emissions, and absolute grid exchange, respectively. In all three maps, a consistent operating region - highlighted by a black rectangle - can be identified around β ≈ 0.5-1.0 and γ ≈ 0.1-0.2. Within this region, cost escalation is limited while both emissions and grid reliance are noticeably reduced, confirming the robustness of this parameter range.

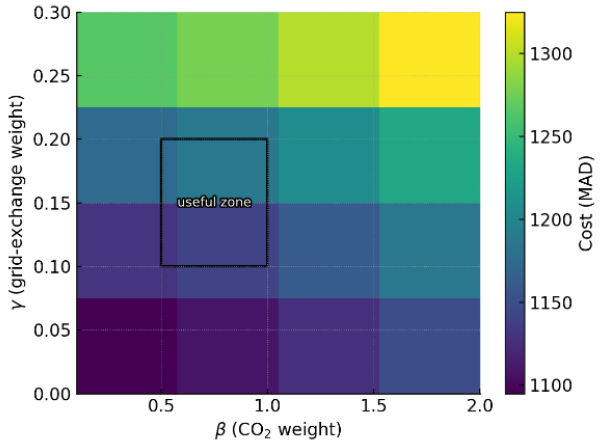



**Fig. 16.** Cost (MAD) over a β×γ sweep.
The marked rectangle highlights a useful operating zone (β ≈ 0.5–1.0, γ ≈ 0.1–0.2) where costs remain moderate while import reliance is curtailed.

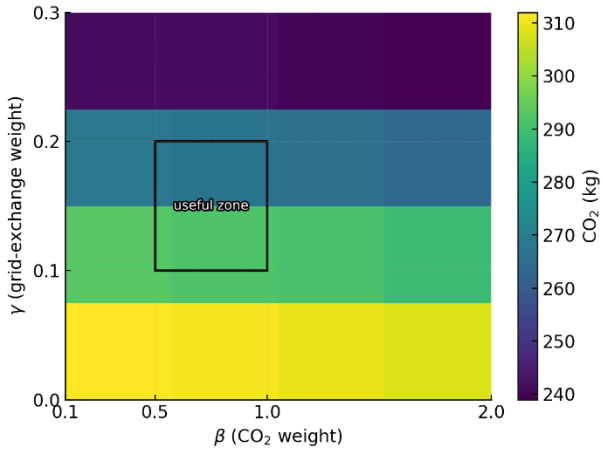



**Fig. 17.** $CO_2$ emissions (kg) across the β×γ sweep. The annotated "useful zone" (β ≈ 0.5–1.0, γ ≈ 0.1–0.2) achieves lower emissions while maintaining improved autonomy with limited cost penalties.

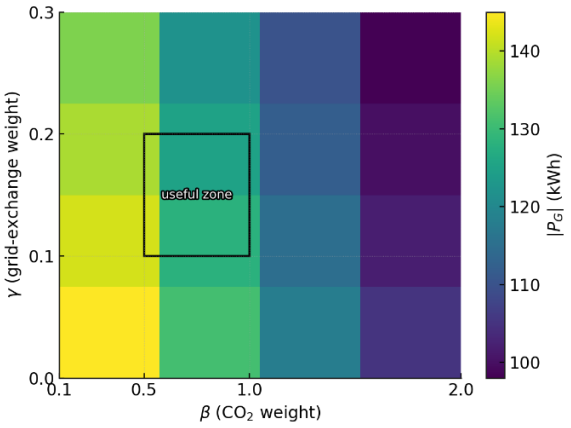



**Fig. 18.** Absolute grid exchange $|P_G|$ (kWh) across the β×γ sweep. Reduced grid dependency is observed in the same operating region, confirming the consistency of the identified trade-off.

Taken together, these results demonstrate that careful tuning of γ and δ enables the controller to navigate smoothly between competing objectives. The radar plot in panel (d) synthesizes these trade-offs by contrasting the three Pareto-efficient operating points of Table 2 - carbon-lean, autonomy-lean, and quality-lean - highlighting the flexibility of the proposed framework to accommodate different operational priorities without compromising feasibility or stability.

### *6.7. Comparative evaluation*

#### *6.7.1. DQN vs PSO (same microgrid)*

Table 3 compares the proposed DQN against the previously implemented PSO [11]. DQN reduces $CO_2$ from 318.9 kg to 272.5 kg (− 14.6 %), and cost from 1250 to 1120 MAD (− 10.4 %). Control latency per episode improves from 12.8 s to 2.1 s (− 83.6 %), enabling edge deployment. Voltage compliance is consistently better under DQN (no distant-node violations).

**Table 3.** DQN vs PSO

| Criteria | DQN-Based Control | PSO-Based Control | %Δ vs PSO |
|---|---|---|---|
| Control strategy type | Model-free, learning-based | Metaheuristic, population-based | - |
| Action space | Discrete, learned policies | Continuous, search-based | - |
| Adaptability to real-time variations | High (real-time reactive) | Low (offline only) | - |
| $CO_2$ emissions (kg) | 272.5 | 318.9 | −14.6% |
| Total cost (MAD) | 1120 | 1250 | −10.4% |
| Voltage compliance | All nodes within ±10% | Occasional violations at distant nodes | Better |
| Battery SOC range maintained | Yes (30–95%) | Partially maintained | Better |
| Computation time (per episode) | ≈ 2.1 s | ≈ 12.8 s | −83.6% |
| Deployment feasibility (edge devices) | Yes | Limited (requires central computation) | Better |

DQN reduces emissions from 318.9 to 272.5 kg $CO_2$ (−14.6%) and lowers total cost from 1250 to 1120 MAD (−10.4%). Control latency per episode drops from about 12.8 s to 2.1 s (−83.6%), making edge implementation practical. Voltage compliance is consistently better under DQN (no violations at distant nodes). PSO remains a solid planning baseline but struggles in real-time operation (search latency, premature convergence), whereas DQN learns tiered, network-feasible responses that generalize under variability, [6], [16], [19], [20], [27], [28].

#### *6.7.2. DQN-MA vs DQN-SA ablation*

*a) Objective.*

Quantify the benefit of decentralized execution under CTDE versus a single centralized DQN, normalized reward Eq. (26), action discretizations, and all constraints Eqs. (1)-(3), (6), (18)–(20)) strictly identical.

*b) Compared configurations.*

- DQN-SA (centralized): one policy outputs the full joint action $(a_t^{PV1}, a_t^{PV2}, a_t^{PV3}, a_t^{WT}, a_t^{DG}, a_t^{B1}, a_t^{B2}, a_t^{B3}, a_t^{PCC})$ input is a global state (generation/load aggregates, PCC indicators, summarized nodal voltages).
- DQN-MA (CTDE): independent learners (PV×3, Wind, DG, Batt×3) trained centrally with global signals and executed locally; a lightweight PCC coordinator enforces import/export exclusivity, zero-crossing deadband, and dwell times.

*c) Experimental design.*

Same hyperparameters and training budget as Table 4. Episodes of 24 h, summer/winter scenarios, plus unseen stressors: PV −20 %, evening-peak +20 %, PV2 outage (4 h), PCC unavailability (2 h). Report means ± std over 5 seeds.

*d) Key performance indicators (to be reported).*

Total cost (MAD); (ii) $CO_2$ (kg); (iii) grid import/export (kWh); (iv) Vviol (p.u.·h, Eq. 19); (v) PCC toggling rate (zero-crossings/day); (vi) diesel runtime (h); (vii) SOC-window adherence (% steps within $[SOC_{min}, SOC_{max}]$); (viii) battery throughput (kWh/day); (ix) episode latency (s) and compute footprint (MB/model or total).

**Table 4.** Ablation: MA-DQN vs SA-DQN (mean ± std, 5 seeds)

| Metric | SA-DQN (Summer) | MA-DQN (Summer) | SA-DQN (Winter) | MA-DQN (Winter) |
|---|---|---|---|---|
| Total cost (MAD) | 1128 ± 14 | 1115 ± 12 | 1176 ± 16 | 1162 ± 15 |
| $CO_2$ (kg) | 279 ± 6 | 271 ± 5 | 296 ± 7 | 287 ± 6 |
| $\|P_G\|$ (kWh) | 138 ± 12 | 121 ± 10 | 162 ± 14 | 138 ± 12 |
| $V_{viol}$ (p.u.·h) | 0.46 ± 0.06 | 0.31 ± 0.05 | 0.52 ± 0.07 | 0.37 ± 0.06 |
| SOC violation (p.u.·h) | 0.20 ± 0.05 | 0.11 ± 0.03 | 0.24 ± 0.06 | 0.14 ± 0.04 |
| PCC sign toggles (count) | 27 ± 6 | 12 ± 3 | 31 ± 7 | 15 ± 4 |
| Diesel runtime (h) | 14.1 ± 0.8 | 13.0 ± 0.7 | 15.6 ± 0.9 | 14.5 ± 0.8 |
| Battery throughput (kWh/day) | 205 ± 11 | 191 ± 10 | 215 ± 12 | 198 ± 11 |
| Episode latency (s) | 2.2 ± 0.1 | 2.0 ± 0.1 | 2.2 ± 0.1 | 2.0 ± 0.1 |
| Compute footprint (MB) | 14.2 ± 0.6 | 17.8 ± 0.8 | 14.2 ± 0.6 | 17.8 ± 0.8 |

*e) Expected patterns (to be confirmed numerically).*

Network quality: Compared with the single-agent DQN, the multi-agent DQN reduces the voltage-violation metric and decreases PCC toggling, owing to local observability and PCC arbitration with deadband and dwell constraints.

Energy & cost: DQN-MA achieves comparable or lower cost and $CO_2$, often with fewer diesel hours and smoother imports.

Storage behavior: fewer SOC-window violations and more balanced battery throughput (healthier cycling).

Compute & deployment: similar or lower latency; distributed memory footprint compatible with edge deployment.

### 6.8. Robustness and Generalization (unseen stressors)

We test three unseen scenarios without retraining: (i) PV profile depression (cloud cover), (ii) PV2 outage (4 h), (iii) +20 % evening peak. In all cases the agent adapts PCC imports, diesel tiers, and battery modes to keep SOC within bounds and preserve supply; voltage remains compliant except a minor spike on node 31 under the load surge.

The quantitative results under these stress scenarios are summarized in Table 5.

The learned policy maintains service continuity and voltage compliance under significant perturbations, key for weakly interconnected Saharan microgrids [11], [13], [24], [37].

**Table 5.** Robustness under unforeseen operating conditions (your values; unchanged)

| Scenario | Total Cost (MAD) | $CO_2$ Emissions (kg) | Diesel Usage (kWh) | Grid Import (kWh) | SOC Range Maintained | Voltage Violation |
|---|---|---|---|---|---|---|
| Baseline (Normal Day) | 1120 | 272.5 | 340.6 | 91.2 | Yes | None |
| Unusual PV Profile (Cloud Cover) | 1245 | 298.1 | 370.2 | 108.5 | Yes | None |
| Partial PV Outage (PV2 disconnected 4h) | 1290 | 322.4 | 388.7 | 120.1 | Yes | None |
| Unexpected Load Surge (+20% peak) | 1375 | 349.7 | 402.9 | 133.0 | Yes | Minor spike at node 31 |

## 7. Conclusion

We proposed and validated a CTDE multi-agent DQN controller for a Saharan hybrid microgrid, combining independent learners (PV, wind, batteries, diesel) with a PCC-aware coordination policy that enforces import/export exclusivity, a zero-crossing deadband, and dwell-time constraints. The formulation aligns the objective and the reward via common policy weights α(cost), β ($CO_2$), γ (import), and δ (power quality), while embedding network feasibility (±10 % voltage band, SOC envelopes, ramp-rates) directly in learning.

On a 24-h digital twin of Boujdour, the learned policy achieves a renewables-first energy mix (PV 56.4 %, grid 9.2 %) and 272.5 kg $CO_2$ (≈ 275 g/kWh), with voltage compliance across representative nodes. Relative to a PSO baseline, the controller lowers total cost by 10.4 % and $CO_2$ by 14.6 %, while cutting control latency by 83.6 %, making the approach suitable for edge hardware. A systematic policy-weight sensitivity confirms carbon-aware tunability: raising β from 0.1 to 2.0 reduces emissions 22.6 % at a 15.5 % cost increase. Robustness under unseen perturbations (cloud-induced PV depression, a 4-h PV2 outage, +20 % peak load) preserves supply continuity, SOC safety, and voltage limits, with only a minor voltage excursion at the feeder tail.

Implications. The results indicate that grid-interactive, voltage-aware DRL can deliver policy-tunable cost–carbon–

autonomy trade-offs in weakly interconnected microgrids, reducing operator intervention while respecting distribution-level constraints.

Limitations. The present controller uses discrete actuation tiers; continuous set-point refinement is delegated to device-level controllers. Exogenous uncertainty (weather/load) is handled implicitly; explicit probabilistic forecasting is not yet integrated. Battery degradation is approximated by throughput-based costs.

Future work. (i) Extend to continuous-action DRL (DDPG/SAC/PPO) to refine set-points without changing supervisory logic; (ii) couple with probabilistic forecasts and distributionally robust training; (iii) add multi-season and multi-year evaluation with realistic outages; (iv) conduct an ablation versus a single-agent DQN and, where available, a MILP benchmark; (v) integrate explainability (policy saliency, action attribution) and cyber-safety constraints for field deployment; (vi) explore multi-microgrid coordination and policy co-design of α, β, γ, δ with local carbon and reliability targets.

These findings position CTDE multi-agent DQN as a credible, edge-deployable supervisory layer for Saharan microgrids, capable of delivering quantifiable gains in cost, carbon, and power quality under operational uncertainty, while remaining tunable to evolving policy priorities.

**Acronyms**

| Acronym | Definition |
|---|---|
| *CTDE* | Centralized Training, Decentralized Execution |
| *MA* | Multi-Agent |
| *EMS* | Energy Management System |
| *PCC* | Point of Common Coupling |
| *BESS* | Battery Energy Storage System |
| *RES* | Renewable Energy Sources |
| *DER* | Distributed Energy Resources |
| *DG* | Diesel Generator |
| *RL/DRL* | (Deep) Reinforcement Learning |
| *DQN/DDQN* | (Double) Deep Q-Network |
| *DDPG/SAC/PPO* | Actor-critic DRL methods (future continuous-action refinements) |
| *MDP* | Markov Decision Process |
| *UC* | Unit Commitment |
| *KPI* | Key Performance Indicator |
| *LV/MV* | Low Voltage / Medium Voltage |
| *PF* | Power Factor |
| *MAD* | Moroccan Dirham |
| *$CO_2$* | Carbon dioxide |

**Main symbols**

| Symbol | Definition |
|---|---|
| $t$ | Time index (h) |
| $\Delta t$ | Time step (h); default Δt = 1 h unless stated |
| $P_G(t)$ | Net PCC exchange (kW); PG>0 import, PG<0 export |
| $P_{DG}(t)$ | Diesel-generator power (kW) |
| $P_{BP}(t)$ | Battery power (kW); positive for discharge, negative for charge (sign follows manuscript) |
| $SOC(t)$ | State of charge (p.u.) |
| $V_i(t)$ | Voltage at bus i (p.u.) |
| $C_{total}(t)$ | Operating cost at time t (MAD) |
| $E_{CO_2}(t)$ | Emissions at time t (kg) |
| $V_{viol}(t)$ | Voltage-violation severity aggregator (p.u.h) |
| $\alpha, \beta, \gamma, \delta$ | Policy weights for cost / $CO_2$ / grid import / power-quality terms |
| $\rho_V, \rho_{SOC}$ | Internal penalty coefficients (reward-shaping, not policy weights) |
| $\zeta$ | PCC zero-crossing deadband (kW) |
| $T_{imp}^{min}, T_{exp}^{min}$ | Minimum up/down (dwell) times for import/export modes (steps) |
| $\gamma_{RL}$ | RL discount factor (default 0.99) |

**Acknowledgment**

The authors would like to thank the Faculty of Sciences Ain Chock (FSAC), Hassan II University of Casablanca, for providing the academic and research environment that supported this work.

**Author Contributions**

All authors contributed to the study conception and design.

Redouane Mihramane: Conceptualization, Methodology, Software, Validation, Writing – original draft, Visualization, Supervision.

Sidi Salah Ech-Charqaouy: Conceptualization, Methodology, Software, Validation, Writing – original draft, Visualization, Supervision.

Abdelkader Boulezhar: Project administration, Resources, Supervision, Validation.

Nizar Ech-Charqaouy: Data curation, Resources, Writing – review & editing.

Amjad Ech-Charqaouy: Software, Formal analysis, Investigation, Writing – review & editing.

All authors read and approved the final manuscript.

**Conflict of Interests**

The authors declare that they have no competing financial or non-financial interests related to the work reported in this manuscript.

**References**

[1] Q. Liu, K. Kamoto, and X. Liu, "Microgrids-as-a-service for rural electrification in Sub-Saharan Africa," Computers, Materials & Continua, vol. 63, pp. 1249–1261, 2020, doi: 10.32604/cmc.2020.05598.

[2] A. Giedraityte, S. Rimkevicius, M. Marciukaitis, V. Radziukynas, and R. Bakas, "Hybrid renewable energy systems: a review of optimization approaches and future challenges," Applied Sciences, vol. 15, no. 4, pp. 1744, 2025, doi: 10.3390/app15041744.

[3] J. Hirwa, A. Zolan, W. Becker, T. Flamand, and A. Newman, "Optimizing design and dispatch of a resilient renewable energy microgrid for a South African hospital," Applied Energy, vol. 348, pp. 121438, 2023, doi: 10.1016/j.apenergy.2023.121438.

[4] Q. Hassan, S. Algburi, A. Z. Sameen, H. M. Salman, and M. Jaszczur, "A review of hybrid renewable energy systems: solar and wind-powered solutions: challenges, opportunities, and policy implications," Results in Engineering, vol. 20, pp. 101621, 2023, doi: 10.1016/j.rineng.2023.101621.

[5] H. Al-Rawashdeh, O. A. Al-Khashman, J. T. Al Bdour, M. R. Gomaa, H. Rezk, A. Marashli, L. M. Arrfou, and M. Louzazni, "Performance analysis of a hybrid renewable-energy system for green buildings to improve efficiency and reduce GHG emissions with multiple scenarios," Sustainability, vol. 15, no. 9, pp. 7529, 2023, doi: 10.3390/su15097529.

[6] M. Ali, K. Prakash, M. A. Hossain, and H. R. Pota, "Intelligent energy management: evolving developments, current challenges, and research directions for sustainable future," Journal of Cleaner Production, vol. 314, pp. 127904, 2021, doi: 10.1016/j.jclepro.2021.127904.

[7] International Energy Agency, "Global Energy Review 2025: $CO_2$ Emissions," IEA, Paris, France, 2025. https://www.iea.org/reports/global-energy-review-2025/co2-emissions.

[8] R. Thapa-Parajuli, R. Neupane, M. Timsina, B. Pokharel, D. Poudel, M. Maharjan, S. P. KC, and S. Shrestha, "Global $CO_2$ emission reduction disparities after and before COVID-19," Sustainability, vol. 17, no. 14, pp. 6602, 2025, doi: 10.3390/su17146602.

[9] I. T. Halidou, H. O. R. Howlader, M. M. Gamil, M. H. Elkholy, and T. Senjyu, "Optimal power scheduling and techno-economic analysis of a residential microgrid for a remotely located area: A case study for the Sahara Desert of Niger," Energies, vol. 16, no. 8, pp. 3471, 2023, doi: 10.3390/en16083471.

[10] L. Richard, C. Boudinet, S. A. Ranaivoson, J. O. Rabarivao, A. E. Befeno, D. Frey, M.-C. Alvarez-Hérault, B. Raison, and N. Saincy, "Development of a DC microgrid with decentralized production and storage: From the lab to field deployment in Rural Africa," Energies, vol. 15, no. 18, pp. 6727, 2022, doi: 10.3390/en15186727.

[11] R. L. Mihramane, S. S. Ech-Charqaouy, D. Saifaoui, N. Ech-Charqaouy, and A. Ech-Charqaouy, "Management and optimization of a renewable energy hybrid system integrated into a microgrid, considering constraints related to voltage stability and service continuity," IJRER, vol. 15, no. 2, pp. 213–225, 2025, doi: 10.20508/ijrer.v15i2.14781.g9042.

[12] S. S. Ech-Charqaouy, D. Saifaoui, O. Benzohra, and A. Lebsir, "Integration of decentralized generations into the distribution network smart grid downstream of the meter," IJsmartGrid, vol. 4, no. 1, pp. 17–27, 2020, doi: 10.20508/ijsmartgrid.v4i1.92.g78.

[13] S. S. Ech-Charqaouy, D. Saifaoui, O. Benzohra, and A. Lebsir, "Impact of integrating renewable energies into distribution networks on the voltage profile," IJRER, vol. 10, no. 1, pp. 143–154, 2020, doi: 10.20508/ijrer.v10i1.10204.g7848.

[14] E. Sutanto, P. E. Astawa, F. Fahmi, M. I. Hamid, M. Yazid, W. Shalannanda, and M. Aziz, "Lithium-ion battery state-of-charge estimation from the voltage discharge profile using gradient vector and support vector machine," Energies, vol. 16, no. 3, Art. no. 1083, 2023, doi: 10.3390/en16031083.

[15] J. A. Braun, R. Behmann, D. Schmider, and W. G. Bessler, "State of charge and state of health diagnosis of batteries with voltage-controlled models," Journal of Power Sources, vol. 544, Art no. 231828, 2022, doi: 10.1016/j.jpowsour.2022.231828.

[16] L. Yi, P. Langlotz, M. Hussong, M. Glatt, F. Sousa, and J. C. Aurich, "An integrated energy management system using double deep Q-learning and energy storage equipment to reduce energy cost in manufacturing under real-time pricing condition: A case study of a scale-model factory," CIRP Journal of Manufacturing Science and Technology, vol. 38, pp. 844–860, 2022, doi: 10.1016/j.cirpj.2022.07.009.

[17] E. Foruzan, L.-K. Soh, and S. Asgarpoor, "Reinforcement learning approach for optimal distributed energy management in a microgrid," IEEE Transactions on Power Systems, vol. 33, no. 5, pp. 5749–5758, 2018, doi: 10.1109/TPWRS.2018.2823641.

[18] D. Liu, C. Zang, P. Zeng, W. Li, X. Wang, Y. Liu, and S. Wang, "Deep reinforcement learning for real-time economic energy management of microgrid system considering uncertainties," Frontiers in Energy Research, vol. 11, Art no. 1163053, 2023, doi: 10.3389/fenrg.2023.1163053.

[19] T. Levent, P. Preux, E. Le Pennec, J. Badosa, G. Henri, and Y. Bonnassieux, "Energy Management for Microgrids: A Reinforcement Learning Approach," in Proc. IEEE PES, Bucharest, Romania, 2019, pp. 1-5, doi: 10.1109/ISGTEurope.2019.8905538.

[20] J. Terven, "Deep reinforcement learning: A chronological overview and methods," AI, vol. 6, no. 3, Art no. 46, 2025, doi: 10.3390/ai6030046.

[21] U. K., K. Gurunadh, N. K. S. Keerthan, N. K. Musunuru, and A. J., "Energy management system-deep reinforcement learning based dueling DQN (Deep Q-Networks)," in Proc. Asia Pacific Conference on Innovation in Technology (APCIT), Mysore, India, 2024, pp. 1–7, doi: 10.1109/APCIT62007.2024.10673561.

[22] E. O. Arwa and K. Folly, "Reinforcement learning techniques for optimal power control in grid-connected microgrids: A comprehensive review," IEEE Access, vol. 8, pp. 227747–227766, 2020, doi: 10.1109/ACCESS.2020.3038735.

[23] M. Ahmed, C. P. Lim, and S. Nahavandi, "A deep Q-network reinforcement learning-based model for autonomous driving," in Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 2021, pp. 739-744, doi: 10.1109/SMC52423.2021.9658892.

[24] G. Custodio, L. F. Ochoa, T. Alpcan, F. C. L. Trindade, and R. A. de Godoy Rosolen, "Using dueling double Q-learning for voltage regulation in PV-rich distribution networks," in Proc. IEEE Madrid PowerTech, Madrid, Spain, 2021, pp. 1-6, doi: 10.1109/PowerTech46648.2021.9494903.

[25] H. Qiu and F. Liu, "A state representation dueling network for deep reinforcement learning," in Proc. IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA, 2020, pp. 669–674, doi: 10.1109/ICTAI50040.2020.00107.

[26] P. Kiran and K. R. M. Vijaya Chandrakala, "New interactive agent based reinforcement learning approach towards smart generator bidding in electricity market with microgrid integration," Applied Soft Computing, vol. 97, Art no. 106762, 2020, doi: 10.1016/j.asoc.2020.106762.

[27] S. Mischos, E. Dalagdi, and D. Vrakas, "Intelligent energy management systems: A review," Artificial Intelligence Review, vol. 56, pp. 1–40, 2023, doi: 10.1007/s10462-023-10441-3.

[28] D. M. S. Zekrifa, R. D. Prakash, M. Dhanalakshmi, R. Puviarasi, C. S. Sundar Ganesh, J. Samson Isaac, P. Duraipandy, and A. Rajaram, "Optimized controller design for renewable energy systems by using deep reinforcement learning technique," IJRER, vol. 14, no. 1, pp. 101-110, 2024, doi:10.20508/ijrer.v14i1.14273.g8866.

[29] P. I. N. Barbalho, A. L. Moraes, V. A. Lacerda, P. H. A. Barra, R. A. S. Fernandes, and D. V. Coury, "Reinforcement Learning Solutions for Microgrid Control and Management: A Survey," IEEE Access, vol. 13, pp. 1–31, 2025, doi: 10.1109/ACCESS.2025.3546578.

[30] C. Sauer, F. J. D. Lange, M. Thurow, I. Dormuth, and A.-L. Boulesteix, "Statistical parametric simulation studies based on real data," Statistics Methodology, 2025, doi: 10.48550/arXiv.2504.04864.

[31] N. Singh, A. Agarwal, and J. Kumar, "A decentralised power management strategy in a hybrid energy storage supported islanded residential sustainable DC microgrid system," IJETP, vol. 18, no. 1, pp. 1–22, 2022, doi: 10.1504/IJETP.2022.10045782.

[32] A. Iosifidis and A. Tefas, Deep learning for robot perception and cognition. cambridge, MA, USA: Academic Press, 2022.

[33] H. Alavizadeh, H. Alavizadeh, and J. Jang-Jaccard, "Deep Q-learning based reinforcement learning approach for network intrusion detection," Computers, vol. 11, no. 3, Art no. 41, 2022, doi: 10.3390/computers11030041.

[34] J. Kumar, A. Agarwal, and N. Singh, "A Cost-Effective DC microgrid system for resilient power supply to grid-connected societies," International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, vol. 36, no. 5, pp. 1–24, 2023, doi: 10.1002/jnm.3097.

[35] T. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," Sustainable Energy, Grids and Networks, vol. 25, Art no. 100413, 2021, doi: 10.1016/j.segan.2020.100413.

[36] S. Stavrev and D. Ginchev, "Reinforcement learning techniques in optimizing energy systems," Electronics, vol. 13, no. 8, Art no. 1459, 2024, doi: 10.3390/electronics13081459.

[37] H. Song, Y. Liu, J. Zhao, J. Liu, and G. Wu, "Prioritized replay dueling DDQN based grid-edge control of community energy storage system," IEEE Transactions on Smart Grid, vol. 12, no. 6, pp. 4950–4961, 2021, doi: 10.1109/TSG.2021.3099133.

[38] E. Samadi, A. Badri, and R. Ebrahimpour, "Q-learning approach for optimal power dispatch of microgrid," in Proc. 28th Iranian Conference on Electrical Engineering (ICEE), Tabriz, Iran, 2020, pp. 1-5, doi: 10.1109/ICEE50131.2020.9260722.

[39] A. Aktaş and Y. Kirçiçek, Solar Hybrid Systems. Cambridge, MA, USA: Academic Press, 2021.

[40] A. Kumari and S. Tanwar, "Reinforcement learning for multiagent-based residential energy management system," in Proc. IEEE Globecom Workshops (GC Wkshps), Madrid, Spain, 2021, pp. 1–6, doi: 10.1109/GCWkshps52748.2021.9682182.

[41] M. K. Perera, K. T. M. U. Hemapala, and W. D. A. S. Wijayapala, "Developing a reinforcement learning model for energy management of microgrids in python," in Proc. International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2021, pp. 68–73, doi: 10.1109/ICCIKE51210.2021.9410754.