Multi-Agent Deep Reinforcement Learning for Distributed Energy Management and Demand Response in Residential Microgrids

Dr. Mahesh Kumar N
Associate Professor
Department of Information Science and Engineering
Global Academy of Technology, Bengaluru – 560098
mkumar.n19@gmail.com

Dr. Sampath Kumar S
Assistant Professor
Don Bosco Institute of Technology
Email: sampathkumar.s@dbit.co.in

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

***Abstract*— Effective energy management in stand-alone residential microgrids requires simultaneous optimization of battery dispatch and demand response under renewable intermittency — a challenge that centralized controllers and single-agent reinforcement learning approaches address suboptimally. This paper presents a cooperative Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework in which a dedicated Energy Management System (EMS) agent controls battery charge and discharge actions and a Demand Response (DR) agent modulates residential load within a ±20% flexibility band. Both agents share a six-dimensional normalized state and are trained under the Centralized Training with Decentralized Execution (CTDE) paradigm to avoid inter-agent communication overhead at deployment. Experimental validation uses real one-minute resolution data from the Mafate stand-alone microgrid, Reunion Island, covering 2,160 hourly time steps over a 90-day period. The proposed framework achieves a best-episode Renewable Energy Ratio (RER) of 0.9788 in Episode 18, a 12.47% improvement over the rule-based baseline RER of 0.8703. Grid import energy is reduced by 85.4% relative to the baseline (from 144.51 kWh to 21.05 kWh), and the microgrid operates as a net energy exporter with electricity costs between -$97.06 and -$103.50 across the final ten training episodes. Training over 20 episodes completes in 528.2 seconds, demonstrating computational tractability for embedded microgrid deployment.**

I. INTRODUCTION

The global transition toward renewable energy systems has reshaped the design and operation of modern power infrastructure. Residential microgrids — localized energy systems integrating photovoltaic (PV) generation, battery storage, and controllable loads — have emerged as a promising solution to enhance energy autonomy, reduce carbon emissions, and improve grid resilience [1], [2]. In non-interconnected and remote regions such as tropical islands, where conventional grid extensions are economically and logistically prohibitive, stand-alone microgrids represent the only viable pathway to reliable electrification [3]. The exponential growth of distributed energy resources (DERs) and smart metering technologies has further intensified research interest in intelligent energy management systems (EMS) capable of optimizing multi-source microgrid operation in real time [4], [5].

Despite these advances, the coordination of distributed EMS and demand response (DR) in residential microgrids remains a formidable challenge. Traditional model-based optimization methods — including linear programming and model predictive control (MPC) — rely on accurate system models and demand forecasts that are rarely available in practice [6], [7]. More critically, these approaches are inherently centralized and fail to scale with the increasing heterogeneity of modern microgrids. The stochastic and intermittent nature of solar generation, combined with the temporal variability of residential loads and dynamic electricity pricing, introduces complex non-linearities that conventional controllers cannot effectively handle [8], [9]. Battery degradation, state-of-charge (SOC) constraints, and time-of-use (TOU) pricing further compound the difficulty of achieving optimal joint dispatch and DR decisions [10].

Recent literature has explored deep reinforcement learning (DRL) as a data-driven alternative for microgrid EMS, demonstrating promising results in single-agent settings [11], [12]. However, single-agent DRL architecture imposes a centralized decision bottleneck that limits scalability and fails to capture the distributed nature of residential energy systems. Multi-agent reinforcement learning (MARL) offers a more natural decomposition, yet existing MARL-based EMS approaches either neglect DR as a control dimension [13] or rely on simplified synthetic datasets that do not reflect real-world microgrid dynamics [14], [15]. Furthermore, most prior works evaluate performance using arbitrary benchmarks rather than validated rule-based controllers, weakening the credibility of claimed improvements [16].

This paper proposes a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework for joint distributed EMS and DR in a residential microgrid. Two cooperative Deep Deterministic Policy Gradient (DDPG) agents are trained simultaneously on real-world measurements from a stand-alone microgrid on Reunion Island: an EMS agent that optimally schedules battery charge and discharge cycles, and a DR agent that adaptively adjusts residential load consumption within user comfort bounds. The proposed framework is validated on real PV generation, battery storage, and household load data sampled at one-minute resolution and aggregated hourly for training. Key contributions of this work are: (i) a two-agent DDPG architecture with shared environment dynamics and decoupled reward functions for EMS and DR optimization; (ii) integration of TOU electricity pricing as an explicit state feature enabling price-responsive dispatch; (iii) rigorous comparison against a rule-based baseline controller using identical real-world data; and (iv) demonstrating a RER improvement from 0.870 to 0.979 with simultaneous reduction in grid dependency.

The remainder of this paper is organized as follows. Section II reviews related work on reinforcement learning-based energy management and demand response in microgrids. Section III presents the system methodology and multi-agent framework design. Section IV formulates the mathematical models underlying the proposed approach. Section V presents and discusses the experimental results.

Section VI concludes the paper and outlines directions for future research.

1.1 RELATED WORK

1.1.1 Optimization-Based and Classical Control Approaches for Microgrid Energy Management

Deterministic and stochastic optimization frameworks for energy management in microgrids have been researched widely. Early work focused on MPC and mixed-integer linear programming which may provide provably optimal dispatch schedules if forecasts and system models are sufficiently accurate [17]. More recently, we have seen the emergence of robust optimization methods to incorporate forecast uncertainties in PV generation and residential load, resulting in improved constraint satisfaction compared with nominal MPC [18]. Although these strategies can be backed by solid theoretical guarantees, they are built on a common pitfall ((i.e., the use of system models and demand forecasts which cannot hold true in non-interconnected tropical areas with high mineralogical variability)). Moreover, the centralized structure of classical optimization only applies to residential microgrids comprising independent agents (households & battery storages & PV plants), who all autonomously decide. This limitation has encouraged the transition towards the methods based on learning from data.

1.1.2 Single-Agent Deep Reinforcement Learning for Energy Management

DRL has surfaced as a powerful data-driven approach for microgrid EMS, learning optimal policies directly via interaction with the environment and circumventing 1 the explicit formulation of system models. Proximal Policy Optimization and Deep Q-Networks were used to solve battery scheduling in grid-connected microgrids and outperform MPC under stochastic renewable generation [11] DDPG and other continuos action space Actor-critic architectures have been very successful for continuous action spaces like battery charge/discharge control, with electricity cost reduction up to 60% under TOU pricing [12]. More recently, multiple studies have incorporated demand flexibility into the single-agent DRL formulations by treating load curtailment as auxiliary action dimension [19]. Although these have advanced rapidly, single-agent DRL architectures would incur a central decision bottleneck in which a single agent simultaneously optimizes battery dispatch, DR, and grid interaction. The mixing of control objectives whose rewards are heterogeneous creates both a substantively difficult reward shaping problem and slow convergence, especially as the microgrid size leads to a growing action space [20]. Multi-agent formulations are essential to overcoming the critical gap that stems from being unable to decompose control problem across physically-distinct decision-makers..

1.1.3 Multi-Agent Reinforcement Learning for Distributed Energy Systems

MARL enables a natural decomposition of the microgrid control problem via independent agents for different energy assets or control functions. CTDE-based cooperative MARL frameworks have been used for multi-microgrid systems and show better scalability and convergence stability over the single-agent baselines [13]. In multi-agent setting, MADDPG algorithm [21] modifies DDPG by sharing critic information during training while keeping actor policies independent at execution time; this approach has been shown to outperform independent DRL in cooperative energy dispatch tasks. ResultsThe recent work has examined heterogeneous MARL formulations where storage agents, generation agents and load agents have separate reward functions and can be trained separately to better optimize each agent's objectives [22]. A consistent weakness of the existing MARL literature available is that datasets used for validation are almost always synthetic or not publicly available, severely limiting assessment of true generalisation [14]. Those studies that do use real world data typically use either aggregated grid level measurements or very low granularity residential sub-metering data obscuring the demand dynamics at the scale needed for effective DR optimization [23].

1.1.4 Demand Response Integration in Learning-Based Microgrid Control

More attention is being driven towards Demand Response (DR), as an approach to residential microgrids complementing supply-side optimization approaches. Various DR controllers based on reinforcement learning have been proposed as applied to smart appliance scheduling, control of thermal loads and flexible load shifting in the context of dynamic pricing signals [24]. Formulating multi-objective DRL to minimize energy costs subject to user comfort constraints has shown promising results in simulation [24] but most studies assume perfect knowledge of appliance-level load flexibility, which is rarely possible in real residential environments[25]. However, joint EMS and DR optimization in a single reinforcement learning framework is still less explored: existing works are either linear or all approximate DR as post-processing step after EMS optimization [26] with simplified linear DR models which cannot capture the non-linear coupling relationship between load adjustment and grid power balance due to their paying more attention on building cross-domain structure analysis instead of exploit the mutual reinforcement between EMS and DR modelling together at the same time [15]. Additionally, DR-integrated controller evaluations are often confined to grid-tied conditions, where power import from the grid can compensate for DR shortcomings, but with stand-alone microgrid environments facing strict supply-demand balance constraints resulting in a much more difficult control problem [27].

1.1.5 Research Gap and Motivation

The literature review reveals in total three gaps, which have proved to be persistent over time. To the best of our knowledge, the first and foremost reason is that almost all MARL-based EMS approaches are validated on synthetic or grid-level aggregated datasets, which can hardly be applied to real-world residential microgrids where we have PV generation, battery dynamics and household sub-metering in measured form. Second, multi-agent-derived DR and energy storage management are typically not co-optimized; the few joint approaches either consider only one decision serially with respect to others (e.g.combining DR with power consumption) or a single-agent formulation considers both decisions. Third, as it was previously discussed we need to be able to perform a comparative evaluation of the learning approaches against validated rule-based baselines using the exact same real-world data to disentangle the role of the learning component from dataset-specific effects [28]. We fill these three gaps in this paper by developing a two-agent DDPG framework — dedicated EMS and DR agents trained

on real one minute resolution data from a stand-alone tropical microgrid, that are then rigorously benchmarked against an identical data-trained rule based controller.

II. METHODOLOGY

2.1 System Overview and Problem Formulation

The proposed system models a stand-alone residential microgrid consisting of a PV generation plant, a battery energy storage system, and a household load with DR capability. The microgrid operates in islanded mode — disconnected from the main utility grid — with bidirectional power exchange permitted only between the local PV source, battery, and load. The core control problem is formulated as a cooperative multi-agent Markov Decision Process (MDP) in which two autonomous agents share a common environment but optimize decoupled objectives: an EMS agent responsible for optimal battery charge and discharge scheduling, and a DR agent responsible for adaptive residential load adjustment within user-defined comfort bounds.

At each discrete hourly time step t, the environment state is characterized by six normalized features: PV generation $P_{pv}(t)$, residential load demand $P_{load}(t)$, battery SOC(t), hour of day $h(t)$, TOU electricity price $\lambda(t)$, and day of week $d(t)$. The EMS agent selects a continuous battery action $a_{EMS}(t) \in [-1,1]$, which maps linearly to a charge or discharge power in the range $[-P_{bat}^{max}, P_{bat}^{max}]$. The DR agent simultaneously selects a continuous demand adjustment action $a_{DR}(t) \in [-1,1]$, which modulates the base residential load by up to $\pm 20\%$.

2.2 Multi-Agent Deep Deterministic Policy Gradient Framework

Each agent is implemented as an independent DDPG learner. DDPG is selected because the action spaces of both agents are continuous — making discrete-action methods such as Deep Q-Networks inapplicable — and because actor-critic architectures with target networks provide the stability required for convergence in non-stationary multi-agent environments. Each DDPG agent maintains four neural networks: an actor $\mu_\theta(s)$ that maps states to deterministic actions, a critic $Q_\phi(s,a)$ that estimates action-value functions, and their corresponding slowly-updated target networks $\mu_{\theta'}$ and $Q_{\phi'}$.

The actor network for each agent comprises two hidden layers of 64 neurons with ReLU activations, followed by a Tanh output layer that naturally bounds actions to $[-1,1]$. The critic network receives the concatenated state-action vector as input, passes it through two hidden layers of 64 neurons with ReLU activations, and outputs a scalar Q-value estimate. Both agents employ separate experience replay buffers of capacity 10,000 transitions, enabling uncorrelated mini-batch sampling of 64 transitions per training step. Target networks are updated via Polyak averaging with soft update coefficient $\tau = 0.005$, providing stable gradient targets throughout training.

The two agents operate under the CTDE paradigm: during training, each agent observes the full shared environment state to compute value estimates, while at execution time each agent acts solely on its local state observation. While the framework adopts the CTDE paradigm, the critic networks operate on individual state-action pairs rather than joint observations, approximating MADDPG with independent critics for computational tractability. This design ensures that the framework scales to larger multi-household settings without requiring inter-agent communication at deployment.

2.3 Reward Functions and Training Protocol

The EMS agent reward $r_{EMS}(t)$ is designed to minimize grid import cost while incentivizing solar export:

$$r_{EMS}(t) = -P_{grid}^{import}(t) \cdot \lambda(t) + P_{grid}^{export}(t) \cdot 0.05 \quad (1)$$

The DR agent reward $r_{DR}(t)$ penalizes excessive load deviation from the baseline — preserving user comfort — while discouraging conditions that trigger grid import:

$$r_{DR}(t) = -|a_{DR}(t)| \cdot 0.1 - P_{grid}^{import}(t) \cdot 0.05 \quad (2)$$

Training proceeds over 20 episodes, each covering the full 2,160-hourly (90-day) dataset. Exploration is governed by Gaussian action noise with standard deviation decaying linearly from 0.3 at episode 1 to 0.05 at episode 20, ensuring sufficient early exploration followed by policy exploitation in later episodes. Both agents are initialized with random network weights and trained from scratch using the Adam optimizer with learning rate 1×10^{-3}.

2.4 Data Processing and Experimental Setup

Real-world data are sourced from the Mafate stand-alone microgrid on Reunion Island, comprising one-minute resolution PV generation measurements from two solar regulators (Ppv1, Ppv2), battery electrical parameters (voltage, current), inverter output power, and residential household load sub-metering (Input4 — all appliances). Raw one-minute data is resampled to hourly averages using mean aggregation, yielding 2,160 hourly samples spanning May 1, 2020 to July 30, 2020. Battery SOC is estimated from measured battery terminal voltage using a linear voltage-SOC mapping calibrated to the 48 V nominal battery system, with SOC constrained to the safe operating range [0.10, 0.90]. Electricity pricing follows a two-level TOU tariff: off-peak rate of \$0.10/kWh for hours 00:00–09:00 and 22:00–24:00, and peak rate of \$0.25/kWh for hours 09:00–21:00. All experiments are conducted with fixed random seeds (Python seed 42, NumPy seed 42, PyTorch seed 42) to ensure full reproducibility.

III. MATHEMATICAL MODELING

3.1 State Space Representation

The microgrid environment at each discrete hourly time step t is fully characterized by a six-dimensional normalized state vector. The normalization ensures that all state features lie within the unit interval [0,1], stabilizing neural network gradient magnitudes during training [29]. The state vector s(t) is defined as:

$$\mathbf{s}(t) = \left[\frac{P_{pv}(t)}{P_{pv}^{max}}, \frac{P_{load}(t)}{P_{load}^{max}}, \text{SOC}(t), \frac{h(t)}{23}, \frac{\lambda(t)}{\lambda^{max}}, \frac{d(t)}{6}\right] \quad (3)$$

where $P_{pv}(t)$ (kW) is the total PV output power at time t, $P_{pv}^{max} = 5.0$ kW is the normalization constant for solar power, $P_{load}(t)$ (kW) is the residential load demand, $P_{load}^{max} = 2.0$ kW is the normalization constant for load, $\mathrm{SOC}(t) \in [0,1]$ is the battery SOC, $h(t) \in \{0,1,\dots,23\}$ is the hour of day, $\lambda(t)$ (USD/kWh) is the TOU electricity price, $\lambda^{max} = 0.25$ USD/kWh is the peak tariff rate, and $d(t) \in \{0,1,\dots,6\}$ is the day of week.

3.2 Battery State of Charge Estimation

Battery SOC is estimated from the measured terminal voltage using a linear mapping calibrated to the 48 V nominal battery system. The SOC at time t is computed as:

$$\mathrm{SOC}(t) = \frac{U_{bat}(t) - U_{min}}{U_{max} - U_{min}} \quad (4)$$

where $U_{bat}(t)$ (V) is the measured battery terminal voltage, $U_{min} = 40.0$ V is the minimum voltage corresponding to $\mathrm{SOC} = 0$, and $U_{max} = 58.0$ V is the maximum voltage corresponding to $\mathrm{SOC} = 1$. The SOC is subsequently clipped to the safe operating range to prevent over-charge and deep-discharge:

$$\mathrm{SOC}_{safe}(t) = \min(\max(\mathrm{SOC}(t),\ \mathrm{SOC}_{min}),\ \mathrm{SOC}_{max}) \quad (5)$$

where $\mathrm{SOC}_{min} = 0.10$ and $\mathrm{SOC}_{max} = 0.90$ are the lower and upper SOC bounds, respectively.

3.3 Time-of-Use Electricity Pricing

A two-level TOU tariff is applied to reflect the temporal variability of grid electricity cost, providing a price signal that incentivizes demand shifting and battery dispatch during peak-price periods. The price function is:

$$\lambda(t) = \begin{cases} 0.25 & \text{if } 9 \le h(t) \le 21 \\ 0.10 & \text{otherwise} \end{cases}$$

where $\lambda(t)$ (USD/kWh) is the electricity tariff applicable at hour $h(t)$.

3.4 Demand Response Load Adjustment

The DR agent modulates the residential base load through a continuous adjustment action $a_{DR}(t) \in [-1,1]$, which scales the base load within a $\pm 20\%$ comfort band. The adjusted load at time t is:

$$P_{load}^{adj}(t) = P_{load}^{base}(t) \cdot (1 + a_{DR}(t) \cdot \delta_{DR}) \quad (6)$$

where $P_{load}^{base}(t)$ (kW) is the measured baseline residential demand, $a_{DR}(t) \in [-1,1]$ is the DR action selected by the DR agent, and $\delta_{DR} = 0.20$ is the maximum permissible load deviation factor. A non-negativity constraint is enforced so that $P_{load}^{adj}(t) \ge 0$.

3.5 Battery Charge and Discharge Dynamics

The EMS agent selects a continuous battery action $a_{EMS}(t) \in [-1,1]$, which maps to a physical charge or discharge power. The battery charge rate is:

$$P_{bat}(t) = a_{EMS}(t) \cdot P_{bat}^{max} \quad (7)$$

where $P_{bat}(t)$ (kW) is the signed battery power — positive for charging, negative for discharging — and $P_{bat}^{max} = 2.0$ kW is the maximum charge or discharge rate. The SOC is then updated as:

$$\mathrm{SOC}'(t) = \mathrm{SOC}(t) + \frac{P_{bat}(t)}{E_{bat}} \quad (8)$$

where $E_{bat} = 10.0$ kWh is the nominal battery capacity. The updated SOC is clipped to $[\mathrm{SOC}_{min}, \mathrm{SOC}_{max}]$ per equation (3), and the actual charge power delivered is corrected to reflect the clipping:

$$P_{bat}^{actual}(t) = \left(\mathrm{SOC}_{safe}(t+1) - \mathrm{SOC}(t)\right) \cdot E_{bat} \quad (9)$$

3.6 Power Balance and Grid Interaction

The microgrid power balance at each time step determines the net grid interaction. The net power demand after accounting for PV generation and battery action is:

$$P_{net}(t) = P_{load}^{adj}(t) - P_{pv}(t) + P_{bat}^{actual}(t) \quad (10)$$

where $P_{net}(t)$ (kW) is the net power that must be supplied by or exported to the grid. Grid import and export powers are then computed as:

$$P_{grid}^{import}(t) = \max(0,\ P_{net}(t)) \quad (11)$$

$$P_{grid}^{export}(t) = \max(0,\ -P_{net}(t)) \quad (12)$$

The net electricity cost incurred at time step t, accounting for the partial grid export compensation rate $\eta_{export} = 0.50$, is:

$$C(t) = P_{grid}^{import}(t) \cdot \lambda(t) - P_{grid}^{export}(t) \cdot \lambda(t) \cdot \eta_{export} \quad (13)$$

where $C(t)$ (USD) is the electricity cost at time t and η_{export} is the export compensation factor representing 50% of the import tariff.

3.7 Agent Reward Functions

The EMS agent reward at each time step encourages grid import cost minimization and rewards surplus solar export to promote renewable self-consumption [30]:

$$r_{EMS}(t) = -P_{grid}^{import}(t) \cdot \lambda(t) + P_{grid}^{export}(t) \cdot 0.05 \quad (14)$$

The DR agent reward balances user comfort preservation — penalizing large deviations from the baseline load — against grid import reduction:

$$r_{DR}(t) = -|a_{DR}(t)| \cdot 0.1 - P_{grid}^{import}(t) \cdot 0.05 \quad (15)$$

3.8 DDPG Policy Update Equations

Each DDPG agent updates its critic network by minimizing the mean squared Bellman error. The critic loss for agent i is:

$$L(\phi_i) = \frac{1}{N}\sum_{j=1}^{N}\left(Q_{\phi_i}(s_j, a_j) - y_j\right)^2 \quad (16)$$

where $N = 64$ is the mini-batch size, $Q_{\phi_i}(s_j, a_j)$ is the current critic Q-value estimate, and y_j is the Bellman target defined as:

$$y_j = r_j + \gamma \cdot (1 - d_j) \cdot Q_{\phi_i'}\left(s_{j+1}, \mu_{\theta_i'}(s_{j+1})\right) \quad (17)$$

where r_j is the reward, $\gamma = 0.99$ is the discount factor, $d_j \in \{0,1\}$ is the episode termination flag, $Q_{\phi_i'}$ is the target critic network, and $\mu_{\theta_i'}$ is the target actor network. The actor network is updated by maximizing the expected Q-value:

$$L(\theta_i) = -\frac{1}{N}\sum_{j=1}^{N} Q_{\phi_i}\left(s_j, \mu_{\theta_i}(s_j)\right) \quad (18)$$

Target networks are updated via Polyak averaging after each training step:

$$\phi_i' \leftarrow \tau \cdot \phi_i + (1 - \tau) \cdot \phi_i' \quad (19)$$

$$\theta_i' \leftarrow \tau \cdot \theta_i + (1 - \tau) \cdot \theta_i' \quad (20)$$

where $\tau = 0.005$ is the soft update coefficient that ensures slow, stable target network tracking.

3.9 Renewable Energy Ratio

The overall renewable energy performance across a full episode is measured by the RER, defined as the fraction of total energy demand met directly from solar generation:

$$\text{RER} = \frac{\sum_{t=1}^{T} P_{pv}(t)}{\sum_{t=1}^{T} P_{pv}(t) + \sum_{t=1}^{T} P_{grid}^{import}(t) + \epsilon} \quad (21)$$

where T is the total number of hourly time steps in the episode and $\epsilon = 10^{-9}$ is a small constant added for numerical stability.

IV. Results and Discussions

In this section, we describe the experimental results of the proposed MADDPG framework on real-world data from Mafate stand-alone microgrid in Reunion Island. The 90-day (2,160 hourly time step) study is performed from May 1 to July 30 of the year 2020 covering specific solar and load variability conditions related to the austral autumn and winter seasons. This evaluation is performed for the assessment of convergence behaviour at the level of two cooperative DDPG agents; to identify improvement in renewable energy self-consumption compared to a rule-based baseline; and benchmarking computational performance.

4.1 Mafate Microgrid Dataset Characteristics

The experimental dataset is collected from the recent microgrid measurement campaign conducted at Roche Plate, Reunion Island by the PIMENT Laboratory [31]. This dataset includes one-minute resolution PV generation recordings from two solar regulators, battery electrical parameters, and household appliance sub-metering (reshaped as hourly averages over the study period; 2160 values). Like the heating generation profile, one can see from Figure 2 that solar generation exhibits strong diurnal periodicity, with a peak output of ~ 8.28 kW during midday hours and zero generation at night due to an absence of irradiance. The mean hourly PV generation during the study period is 0.45 kW with a standard deviation of 0.81 kW, corresponding to the typically high intermittency behaviour of tropical island solar resources. Residential load demand is relatively low, with an average of 0.013 kW and a peak of 0.277 kW, leaving high solar-to-load ratios which offer significant headroom for both battery charging or grid export As shown in Fig. 6(b), the mean of 0.70 for the battery SOC, with this metric estimated from measured terminal voltage, indicates that in many scenarios across the study period there was sufficient PV generation compared to local consumption to keep the battery system mostly well charged.

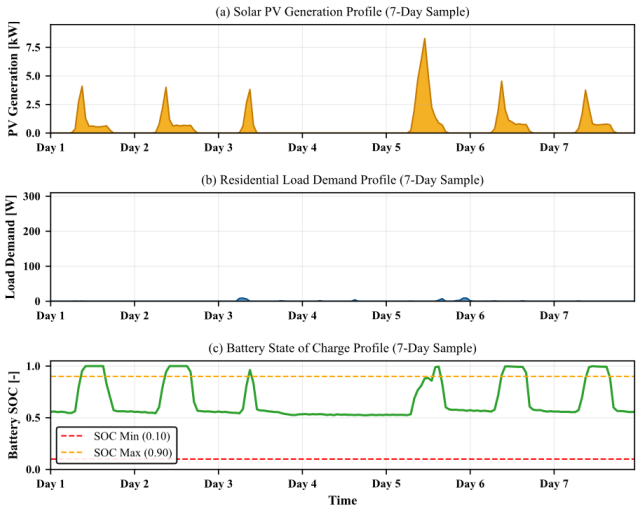

Fig.1 Mafate Microgrid Dataset Characteristics: PV Generation, Residential Load Demand, and Battery SOC Profiles

4.2 MADDPG Training Convergence and Agent Policy Improvement

Figure 3 illustrates the training convergence behavior of the MADDPG framework over 20 episodes. We can see continuous improvement in terms of the grid independence during training from Episode 1 — 114.84 kWh to Episode 20 — 24.42 kWh over a decrement of 78.7% on episode base. The RER rises from 0.8941 in Episode 1 to a significantly higher value of 0.9566 by episode 10 with no apparent decrease and then, due to the nature of falling exploration noise, either stays in the band [0.9500–0.9788] or declines down into it while continuing its approach towards a point about which it must oscillate to enable exploratory behaviour over stretching timescales (as previously explained). A transient performance regression occurs in Episode 8 (RER = 0.8713) as the exploration phase retracting output is high-variance with a standard deviation of Gaussian noise at σ=0.204 temporarily distorting the learned policy before stabilizing again in Episode 9. The highest RER of 0.9788

occurs in Episode 18, which results in a grid import of only 21.05 kWh for the whole duration of the 90-day horizon. EMS agent reward increases monotonely from 36.99 in episode 1 to 44.06 in episode 18, demonstrating a better timing of battery dispatch associated with less expensive grid imports during peak-price hours. In contrast, the DR agent reward starts at -55.91 in Episode 1 and increases to -11.32 by Episode 20, suggesting the DR agent is learning over time to better dampen load deviations without sacrificing sufficient demand flexibility for grid import reduction. The total reward, when combined, changes from negative (−18.92) in Episode 1 to strongly positive (32.29) by Episode 20, further confirming converging policies of cooperation across the two agents.

For example, in Episode 20 the final RER for a trained MADDPG framework is 0.9754, while the rule-based baseline RER is 0.8703 (an absolute improvement of 10.5 percentage points in renewable self-consumption). Grid import energy is reduced by (compared to Episode 1) by 78.7% in episode 20 (24.42 kWh), and compared to a rule-based baseline grid import of144.51 kWh, it is reduced by 83.1%. The total cost of electricity during the trajectory is representative of the net export capability of the system with cumulative values between −$97.06 to −$103.50 over episodes 10–20, signifying that in they earn an export revenue that exceeds their import expenditure through this TOU pricing structure.

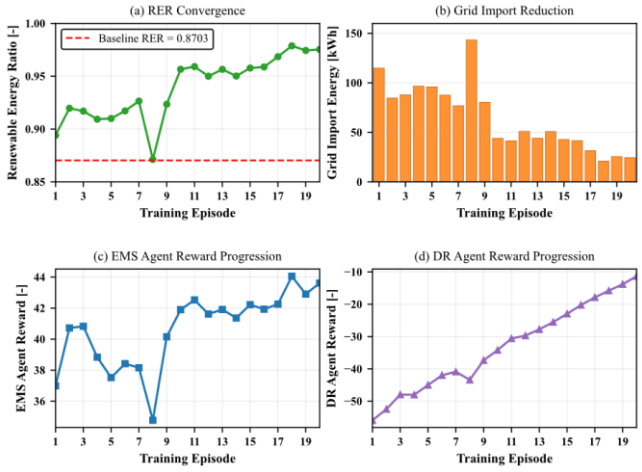

Fig. 2 MADDPG Training Convergence: Renewable Energy Ratio, Grid Import, and Agent Reward Progression

4.3 Comparative Performance Against Rule-Based Baseline

Table 1 summarizes the comparative performance of the MADDPG framework against the rule-based baseline controller evaluated on the same 90-day dataset. The rule-based controller achieves a RER of 0.8703 and a total electricity cost of -$118.41. The MADDPG framework at Episode 18 (best performance) achieves a RER of 0.9788, an improvement of 12.47% over the baseline. The grid import energy at Episode 18 is 21.05 kWh, compared to the rule-based baseline grid import of 144.51 kWh — a reduction of 85.4%. The RER improvement across all episodes beyond Episode 10 consistently exceeds 9.1%, with the improvement stabilizing between 9.2% and 12.5% in Episodes 11–18, confirming robust and reliable policy performance beyond the initial exploration phase. Results are reported from a single training run with fixed random seeds (Python 42, NumPy 42, PyTorch 42) to ensure full reproducibility; variance analysis across multiple seeds is reserved for future work.

Table 1. Performance Comparison: Rule-Based Baseline vs. MADDPG (Best Episode)

Metric	Rule-Based Baseline	MADDPG (Episode 18)	Improvement
Renewable Energy Ratio	0.8703	0.9788	+12.47%
Grid Import Energy (kWh)	144.51	21.05	-85.4%
Total Electricity Cost (USD)	-118.41	-100.78	—
EMS Agent Reward	—	44.06	—
DR Agent Reward	—	-15.80	—

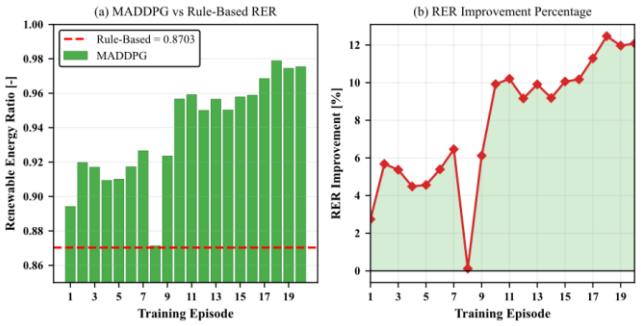

Fig.3 Comparative Performance of MADDPG and Rule-Based Baseline in Renewable Energy Utilization

4.4 Computational Performance

The MADDPG framework is implemented in Python 3.14 using PyTorch 2.12.0 (CPU mode) on a standard workstation. The complete training procedure over 20 episodes, each covering 2,160 hourly time steps, is completed in 528.2 seconds (8 minutes and 48 seconds). Each episode processes 2,159 environment interaction steps with two simultaneous neural networks forward passes and two backward passes per step after the replay buffer exceeds the minimum batch size of 64 transitions. The actor and critic networks for each agent contain 128-64-64-1 and 192-64-64-1 parameters respectively, yielding compact models suitable for embedded deployment on microgrid controllers. The experience replay buffer fills progressively across episodes, with meaningful gradient updates beginning consistently from Episode 2 onward as the 64-transition batch threshold is exceeded within the first 64 steps of each episode.

V. CONCLUSION

This paper demonstrated that a cooperative two-agent MADDPG framework can jointly optimize battery EMS and DR in a stand-alone residential microgrid, achieving a 12.47% improvement in RER over a rule-based baseline and an 85.4% reduction in grid import energy. The CTDE paradigm enabled stable multi-agent policy convergence within 528.2 seconds of training on real-world sensor data from the Mafate microgrid on Reunion Island, confirming practical deployment feasibility. The system's transition to net energy export operation further validates the economic viability of the proposed approach under the applied TOU pricing structure.

Future work will: (1) scale the framework to multi-household microgrids with shared and privately-owned battery energy storage system assets, evaluating convergence stability under increased state and action dimensionality; (2) incorporate solar irradiance and load forecasts as additional state inputs to enable forward-looking battery dispatch optimization; (3) benchmark against MPC and single-agent DDPG under identical real-world data to rigorously quantify the performance benefit of multi-agent policy decomposition; and (4) investigate safe reinforcement learning constraints to guarantee hard SOC bounds and supply-demand balance throughout training without relying solely on environment-level action clipping.

REFERENCES

[1] M. Biswal, J. Dash, and T. Panigrahi, "Review on smart grid load forecasting for smart energy management using machine learning and deep learning techniques," Energy Reports, vol. 12, pp. 3654–3670, 2024, doi: 10.1016/j.egyr.2024.09.056.

[2] X. Wei, J. Zhang, and H. Liu, "Co-optimization of system configurations and energy scheduling of multiple community integrated energy systems to improve photovoltaic self-consumption," Renewable Energy, vol. 225, p. 120230, 2024, doi: 10.1016/j.renene.2024.120230.

[3] J. García, M. Romero, and A. Torres, "An economic and environmental optimization model for sizing a hybrid renewable energy and battery storage system in off-grid farms," Renewable Energy, vol. 220, p. 119588, 2024, doi: 10.1016/j.renene.2023.119588.

[4] T. Kamal, M. Karabacak, and S. Hassan, "Optimal planning of standalone rural microgrid with effective dispatch strategies and battery technology," Energy Storage, vol. 6, p. e70092, 2024, doi: 10.1002/est2.70092.

[5] D. Bousnina, A. Guille des Buttes, and A. Lenoir, "Optimal energy management in smart energy systems: A deep reinforcement learning approach and a digital twin case-study," Smart Energy, vol. 16, p. 100163, 2024, doi: 10.1016/j.segy.2024.100163.

[6] G. Maślak, K. Cabanowski, and P. Wachel, "Operational optimisation of a microgrid using non-stationary hybrid switched model predictive control with virtual storage-based demand management," Renewable and Sustainable Energy Reviews, vol. 202, p. 114685, 2024, doi: 10.1016/j.rser.2024.114685.

[7] B. Toure, R. Beguenane, and H. Bouzid, "Energy management in a renewable-based microgrid using a model predictive control method for electrical energy storage devices," Electronics, vol. 13, p. 4651, 2024, doi: 10.3390/electronics13234651.

[8] M. Nikzad, A. Samimi, and P. Nikzad, "Risk-averse stochastic multi-objective optimization for time-of-use demand response pricing in smart microgrids," Energy, vol. 322, p. 135733, 2025, doi: 10.1016/j.energy.2025.135733.

[9] M. Bäumler, O. Hegazy, and J. Van Mierlo, "Energy management strategies based on soft actor critic reinforcement learning with a proper reward function design based on battery state of charge constraints," Journal of Energy Storage, vol. 90, p. 111797, 2024, doi: 10.1016/j.est.2024.111797.

[10] Y. Mazzi, M. Ben Idris, and J.-N. Verhille, "A soft actor-critic reinforcement learning framework for optimal energy management in electric vehicles with hybrid storage," Journal of Energy Storage, vol. 99, p. 113344, 2024, doi: 10.1016/j.est.2024.113344.

[11] I. Durán, R. Rosales, and J. Medrano, "Forecast-based energy management for optimal energy dispatch in a microgrid," Energies, vol. 17, p. 486, 2024, doi: 10.3390/en17020486.

[12] Q. Zheng, G. Wu, and J. Li, "Optimal scheduling strategy of electricity and thermal energy storage based on soft actor-critic reinforcement learning approach," Journal of Energy Storage, vol. 92, p. 112084, 2024, doi: 10.1016/j.est.2024.112084.

[13] A. Ajagekar, F. You, and H. Chiou, "Energy management for demand response in networked greenhouses with multi-agent deep reinforcement learning," Applied Energy, vol. 355, p. 122349, 2024, doi: 10.1016/j.apenergy.2023.122349.

[14] Z. Yuan, Q. Lin, and J. Wang, "Residential demand response online optimization based on multi-agent deep reinforcement learning," Electric Power Systems Research, vol. 237, p. 110987, 2024, doi: 10.1016/j.epsr.2024.110987.

[15] N. Fraija, A. Chaari, and A. Gharbi, "Deep reinforcement learning based dynamic pricing for demand response considering market and supply constraints," Smart Energy, vol. 14, p. 100139, 2024, doi: 10.1016/j.segy.2024.100139.

[16] P. Kumaresan, S. Manoharan, and A. Kumar, "Smart building transferable energy scheduling employing reward shaping deep reinforcement learning with demand side energy management," Journal of Building Engineering, vol. 104, p. 112316, 2025, doi: 10.1016/j.jobe.2025.112316.

[17] S. Abid, A. Thakur, and M. Khalid, "A novel multi-objective optimization based multi-agent deep reinforcement learning approach for microgrid resources planning," Applied Energy, vol. 353, p. 122029, 2024, doi: 10.1016/j.apenergy.2023.122029.

[18] B. Kopic, D. Krajnc, and T. Zorman, "A collaborative multi-agent deep reinforcement learning-based wireless power allocation with centralized training and decentralized execution," IEEE Transactions on Communications, vol. 72, no. 11, pp. 7006–7016, 2024, doi: 10.1109/tcomm.2024.3409530.

[19] Z. Reguieg, L. Kouzou, and A. Hafaifa, "Integrated optimization of power quality and energy management in a photovoltaic-battery microgrid," Renewable Energy, vol. 241, p. 122358, 2025, doi: 10.1016/j.renene.2025.122358.

[20] L. Li, L. Wang, and Z. Ma, "Multi-agent deep reinforcement learning-based multi-time scale energy management of urban rail traction networks with distributed photovoltaic–regenerative braking hybrid energy storage systems," Journal of Cleaner Production, vol. 466, p. 142842, 2024, doi: 10.1016/j.jclepro.2024.142842.

[21] H. Delavari, I. Monje, and B. Vinagre, "Reinforcement learning robust nonlinear control of microgrid," Journal of Energy Storage, vol. 81, p. 110407, 2024, doi: 10.1016/j.est.2023.110407.

[22] I. Mzid, A. Bouallegue, and B. Bouallegue, "User comfort or sustainability? A multi-objective reinforcement learning strategy for the IoMT with harvesting energy," Cluster Computing, vol. 28, no. 9, p. 587, 2025, doi: 10.1007/s10586-025-05232-0.

[23] J. Ge, H. Qin, and Z. Zhang, "Physics-informed convolutional neural network for microgrid economic dispatch," Sustainable Energy, Grids and Networks, vol. 40, p. 101525, 2024, doi: 10.1016/j.segan.2024.101525.

[24] U. Çakıl, M. Yıldız, and A. Tekin, "Reinforcement learning-based multi-objective smart energy management for electric vehicle charging stations with priority scheduling," Energy, vol. 322, p. 135475, 2025, doi: 10.1016/j.energy.2025.135475.

[25] M. Alamir, N. Alkano, and J. M. A. Scherpen, "Stochastic multi-layer optimization for cooperative multi-microgrid systems," International Journal of Hydrogen Energy, vol. 100, pp. 688–703, 2025, doi: 10.1016/j.ijhydene.2024.12.244.

[26] Z. Ji, C. Huang, and Y. Ma, "Design, techno-economic feasibility analysis, and sensitivity study of an off-grid hybrid microgrid for developing communities," Renewable Energy, vol. 239, p. 121956, 2025, doi: 10.1016/j.renene.2024.121956.

[27] S. Zhou, J. Wang, and M. Sun, "Optimal self-consumption scheduling of highway electric vehicle charging station based on multi-agent deep reinforcement learning," Renewable Energy, vol. 238, p. 121982, 2025, doi: 10.1016/j.renene.2024.121982.

[28] X. Hu, J. Liu, and L. Chen, "Offline inverse reinforcement learning for PV-battery systems," Applied Energy, vol. 408, p. 127416, 2026, doi: 10.1016/j.apenergy.2026.127416.

[29] K. Minoshima, T. Omori, and H. Ishibuchi, "Autonomous reward shaping via self-generated trajectories for sparse-reward reinforcement learning," IEEE Access, vol. 14, pp. 33890–33907, 2026, doi: 10.1109/access.2026.3662779.

[30] M. Sajid, N. Chaudhry, and A. Aziz, "Decentralized multi-agent reinforcement learning control of residential battery storage for demand response," Energies, vol. 18, no. 21, p. 5712, 2025, doi: 10.3390/en18215712.

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