Journal of Operation and Automation in Power Engineering

vol. 13, no. Special Issue, Dec. 2025, Pages: 45-61

http://joape.uma.ac.ir

Research Paper

Decentralized Energy Management in Electrical and Thermal Microgrids Utilizing Reinforcement Learning

Umarov Shukhrat [1, *], *Isaqova Matluba* [2], *Otabek Mukhitdinov* [3], *Boboxujayev Kudrat* [4], *Abdullayev Dadaxon* [5], *Luqmon N Samiev* [6], *Nosirov Nozimbek* [7], and *Sapayev Valisher* [8]

[1]*Department of Engineering of Electrical Machines and Drives, Tashkent State Technical University, University Street No2, Tashkent, Uzbekistan.*
[2]*Tashkent Institute of Irrigation and Agricultural Mechanization Engineers Institute" National Research University, Kari Niyazov Street 39, 100000, Tashkent, Uzbekistan.*
[3]*Kimyo International University in Tashkent, Shota Rustaveli Street 156, 100121, Tashkent, Uzbekistan.*
[4]*PhD, Assistant Professor, Alfraganus University, Uzbekistan.*
[5]*Tashkent Institute of Irrigation and Agricultural Mechanization Engineers National Research University, Uzbekistan.*
[6]*Urgench State University named after Abu Rayhan Biruni, Urgench, Uzbekistan.*
[7]*Research Institute of Environmental and Nature Protection Technologies, 100000, Tashkent, Uzbekistan.*
[8]*Department of General Professional Subjects, Mamun University, Khiva, Uzbekistan.*

Abstract— *This paper proposes a fully decentralized reinforcement learning–based energy management framework for hybrid electrical–thermal microgrids with distributed energy resources. Uncertainties in renewable energy generation, variations in load demand, and the nonlinear nature of battery systems make it difficult to achieve optimal energy management in microgrids. Additionally, using centralized controller techniques in large-scale systems increases computational complexity and makes controller procedure implementation more challenging. This study proposes a fully decentralized multi-agent architecture in which the stochastic performance of agents in the microgrid is modeled using Markov decision processes. This model treats consumers, batteries, and distributed thermal and electrical resources as intelligent, self-governing agents that learn from their surroundings and converge to their best policies through decentralized exploitation. The proposed model-free learning-based approach is designed to not only maximize the profits of producers but also minimize the costs for consumers and reduce the microgrid's reliance on the main grid. Finally, using real-world data from renewable power plants and electricity market data, the performance of the proposed method is evaluated through simulation and accuracy assessment.*

Keywords—Decentralized energy management, microgrid, distributed resources, reinforcement learning, markov decision process.

1. Introduction

The energy industry transition from traditional centralized systems to distributed energy resources (DERs) is fundamentally reshaping the structure of power grids. Distributed energy resources have garnered significant attention due to their environmental benefits and pivotal role in clean and sustainable energy generation [1]. These resources offer the potential to reduce greenhouse gas emissions, transmission power losses, and infrastructure costs. Microgrids, as small-scale and autonomous power networks, provide a suitable platform for leveraging DERs [2]. Distributed energy resources encompass renewable technologies such as wind turbines and solar panels, non-renewable sources like diesel generators, and energy storage systems (ESS) such as batteries [3]. Microgrids can drive in either grid-connected or islanded mode, enhancing grid reliability and providing maintainable, high-quality energy [4].

The use of renewable energy bases is crucial for decreasing need on fossil fuels, lowering greenhouse gas emissions, and ensuring a sustainable and environmentally friendly energy future [5]. Incorporating renewables not only minimizes environmental impact but also fosters economic growth and technological innovation [6]. However, the planning and operation of microgrids present numerous challenges due to uncertainties in load demand forecasting and renewable energy generation [7]. While microgrids offer a flexible pathway for integrating renewable energy sources into power grids, the intermittent nature of these resources introduces challenges in energy management [8]. For instance, photovoltaic (PV) [9] units can only produce electricity in the existence of solar irradiance, and wind farms require sufficient wind speeds to operate effectively [10]. Furthermore, reducing the reliance of microgrids on the main grid is essential for optimizing the profits of producers and minimizing costs for consumers [11]. One effective approach to improve microgrid performance is the use of energy storage systems, such as batteries, which enable energy storage and supply at different times [12]. However, the optimal management of battery charging and discharging, due to nonlinear behavior and dependence on the current state of charge (SOC) and charge/discharge history, becomes a sequential decision-

Received: 27 Nov. 2025
Revised: 23 Dec. 2025
Accepted: 25 Dec. 2025
*Corresponding author:
E-mail: shumarov1951@mail.ru (U. Shukhrat)
DOI: 10.22098/joape.2025.18916.2468

making problem within a dynamic system [13, 14]. Furthermore, the time-varying and state-dependent characteristics of batteries create novel challenges in energy organization planning, requiring the use of more advanced control algorithms [15].

Numerous methods have been established in current years to address uncertainty in microgrid energy management [16]. In [17], a control strategy is presented for the coordinated operation of microgrids within a distribution organization, where the dissemination network operator and each microgrid are considered as independent entities with distinct objective functions to minimize operating costs. This problem is formulated as a two-level stochastic model, with the network operator modeled at the upper level and the microgrids at the lower level. In [18], a stochastic model is presented for microgrid energy scheduling that considers the operational challenges of controllable loads and energy resources. In [19], a robust curved optimization perfect is presented for microgrid energy organization, which reduces the cost of exchanged energy, loads full by energy resources, and storage in grid-connected mode, and minimizes unmet loads in islanded mode, taking into account consumer priorities. In [20], an connected energy organization technique for real-time microgrid process is presented, considering load flow and system performance constraints, where the problem is solved as a stochastic optimal power flow model using Lyapunov optimization. In [21], a two-stage optimization method is presented, where day-ahead hourly scheduling is performed in the first stage, and economic dispatch and real-time energy exchange are scheduled in the second stage using Lyapunov optimization. In [22], two central controllers for the microgrid and gas network are used for energy management in a microgrid, and energy trading is optimized using a mixed-integer linear model in GAMS software. This dependence on estimation models makes the accuracy of these methods a function of the prediction model's accuracy. In contrast, reinforcement learning provides a model-free approach to solving optimal control problems for dynamic systems, which does not require prior information about the stochastic characteristics of the processes.

In [23], a building energy management system was developed to reduce peak energy consumption using a two-stage optimization algorithm. In [24], cost and energy consumption in the presence of renewable resources were optimized using a genetic algorithm. In [25], the impact of demand response in microgrids was investigated using shark smell and grey wolf algorithms, reducing production costs and network losses. Although methods based on heuristic algorithms do not require mathematical modeling and perform better in optimizing nonlinear and discontinuous problems, these methods lack the ability to mimic learning and store prior knowledge [26]. In each step, a new population is randomly selected, which increases the computational time to find the optimal point [27]. In addition, game theory has also been used in the design of microgrid energy management systems. In [28], day-ahead scheduling of microgrids and distribution companies is implemented based on game theory. In this method, demand scheduling and energy storage are modeled as a multi-objective optimization problem. In [29], the output of renewable resources is estimated, and in [30], the energy management problem is solved using game theory techniques. In [31], multi-agent energy management of a microgrid with renewable resources and seasonal loads is implemented based on non-cooperative game theory. In [32], the planning of a power network in grid-connected mode is modeled using cooperative and non-cooperative game theory, in which wind turbines, solar panels, and batteries are considered as players in the problem. The existence of a Nash equilibrium point has been proven through the concavity analysis of return functions and the uncertainty model.

Due to these characteristics, the use of reinforcement learning methods has received widespread consideration in current years. Compared to supervised and unsupervised learning, these methods offer interesting capabilities in the field of control applications. In control systems, due to the difficulty in obtaining initial information and accurate modeling, reinforcement learning is capable of providing model-free approaches to solve problems with uncertainty. These methods improve the performance of the learning agent by storing experiential information. Several studies have used reinforcement learning for energy management. For example, [33] uses a data-driven method based on neural networks and Q-learning for building energy management. In [34], a reinforcement learning-based adaptive dynamic programming technique is developed for smart microgrid control. Also, in [35], decentralized multi-agent energy management for electrical loads of a microgrid is implemented using Q-learning. In [36], hierarchical reinforcement learning is developed to calculate the optimal policy, which converges to a recursive optimal policy. In [37], distributed economic load dispatch is performed using cooperative reinforcement learning based on information received from neighbors and the Diffusion strategy. Q-learning algorithm is an extension of Q-learning for non-cooperative multi-agent systems [38]. In this algorithm, each agent receives not only its own reward but also the rewards and actions of other agents. In practice, access to information about the rewards and actions of other agents is not easily possible for all consumers and producers or even a central system. Combining deep learning and reinforcement learning, known as deep reinforcement learning, has emerged as an approach to solve the problem of the dimensions of Q-functions. In this method, value functions and policies are estimated using deep neural networks. In [39], a deep Q-network is developed to solve problems with a large number of input sensors. In [40], real-time scheduling of microgrids is performed using deep neural networks for function estimation. In [41], in order to intensification the flexibility and reliability of microgrids equipped with renewable resources, a Proximal Policy Optimization algorithm based on deep reinforcement learning and central Critic neural networks is used. With increasing interest in hybrid learning methods, the integration of reinforcement learning and deep learning has emerged as an effective solution. Deep methods help reinforcement learning to solve the dimensionality problem in calculating the Q-function for a large number of agents; however, some of the challenges present in centralized methods remain unresolved. In these methods, information about all agents, including actions and rewards, must be available in a central control unit. Many microgrid energy management methods utilize a centralized control construction. One unit is chosen to be the primary controller in centralized controllers, and it is in charge of overseeing other units. Every agent in this architecture communicates with the central controller. On the other hand, controllers in decentralized structures do not directly communicate with one another, whereas in distributed structures, they only communicate with their neighbors [42]. Centralized control presents significant challenges in large-scale power systems where communications are scarce or unreliable and distributed generation units are scattered throughout the network [43]. For this reason, distributed and decentralized structures have gained attention in multi-agent control systems. Although extensive studies have examined energy management in microgrids, most existing approaches rely on centralized optimization or single-layer control, which becomes impractical in systems containing heterogeneous DERs and tightly coupled electrical–thermal dynamics. Centralized strategies require complete system observability and high computational resources, while fully independent agents ignore subsystem interactions and often converge to suboptimal policies. These limitations underscore the need for a scalable and fully decentralized learning architecture capable of coordinating electrical and thermal subsystems under stochastic demand–generation profiles.

In most previous research, energy management has primarily focused on electrical loads, with less attention paid to thermal loads. Furthermore, energy scheduling in many of these methods has been performed in a centralized or distributed manner, but an approach that simultaneously combines consumption management, optimization of consumer and producer profits, and price offer generation has not been thoroughly investigated. In some studies, battery lifetime has been considered, but the impact of the battery model on the number of replacements has not been taken into account in the calculations. To close this gap, this study proposes a hierarchically coupled, dual-timescale decentralized RL architecture with modified Q-learning policies that encode operational penalties locally. This design enables each agent to learn independent yet coordinated policies without requiring global state information or centralized training. The resulting framework enhances scalability, robustness to uncertainty, and adaptability to hybrid microgrid environments.

Accordingly, in this research, a comprehensive decentralized architecture for the microgrid energy management system is designed, such that agents perform the learning and decision-making process without needing information from their neighbors, and solely based on receiving environmental states. This system is implemented using reinforcement learning and without dependence on the uncertainty model in supply and demand. In this research, the energy management of both electrical and thermal loads is considered, and the microgrid includes distributed electrical and thermal energy resources, a battery energy storing structure, and electrical and thermal consumption loads. The main goal of the system design is to increase the profit of production resources, reduce consumer costs, and reduce the microgrid reliance on the main grid. In addition, by considering battery lifetime, costs resulting from its degradation will also be minimized. The main innovations of this research are as follows:

1) Presentation of a decentralized architecture for a multi-agent energy management system in microgrids, considering distributed electrical and thermal energy resources, a battery energy storing scheme, and thermal and electrical consumption loads.
2) Design of a model-free method based on reinforcement learning for hourly scheduling of the system, without requiring the availability of the uncertainty model in supply and demand.
3) Ability of producer agents to offer prices and decide on the amount of output power (except for renewable energy resources), in order to optimize sales profits and reduce operating costs.
4) Management and reduction of electrical and thermal consumer costs, and decrease of micro grid reliance on main grid.
5) Consideration of the battery lifetime model and minimization of costs resulting from its degradation to increase efficiency and reduce replacement costs.
6) Utilizing actual data from consumers and renewable energy sources to assess the suggested method's accuracy in the electrical grid.

By focusing on a decentralized framework based on reinforcement learning, this research provides a novel approach for the simultaneous management of electrical and thermal loads, which, while reducing reliance on uncertainty models, leads to the optimization of the economic and technical performance of the microgrid.

2. Microgrid Architecture

2.1. Microgrid

A microgrid is a small-scale, low-voltage, and autonomous power network that connects distributed energy resources and loads. These resources include renewable energy, non-renewable energy, and storage systems such as batteries. Microgrids can operate in both grid-connected and islanded modes. In general, it is assumed that microgrids are connected to the main grid. The connection of microgrids to core grid is achieved through the PCC. In grid-connected mode, the microgrid can maintain supply and demand balance by selling surplus energy to the main grid or buying energy in case of a shortage.

One of key objectives in microgrid energy organization is to reduce dependence on the main grid. Therefore, the microgrid energy organization system should be designed in such a way that, while increasing producer profits, grid dependence on core grid is similarly condensed. The loads in micro grids are separated into two groups: controllable and non-controllable:

1) Non-controllable loads include essential consumption such as equipment in medical centers and some industrial processes that must be supplied at the moment of demand. These loads do not have temporal flexibility and cannot be shifted over time.
2) Controllable loads have the ability to be shifted to off-peak hours or even reduce the amount of consumption, which can be effective in optimizing energy management.

Fig. 1 demonstrations the construction of a micro grid that includes solar boards, wind turbines, diesel generators, fuel cells (electrical and thermal), electrical and thermal microturbines, a battery storing structure, and local loads. In this system, the microgrid Operator acts as a high-level controlling agent in the power and energy management of power microgrids.

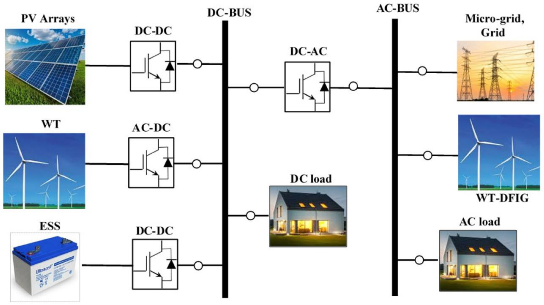


Fig. 1. Microgrid architecture.

2.2. Objective functions

The goal of the energy management system in a micro grid is to exploit profit of all managers within the network over a long period. The overall profit of the i^{th} producer agent over time is defined as follows:

$$\sum_{t} \gamma^{t} \left(P_{i,mic}(t) P_{ri}(t) + P_{i,main}(t) S_{p}(t) - C_{i,op}(t) \right) \quad (1)$$

where t is the time interval of interest and γ is the discount rate. This constraint indicates the present rate of coming rewards; such that the closer its value is to one, the more attention the producer will pay to future profits. $P_{i,mic}(t)$ and $P_{i,main}(t)$ are the power sold from the i^{th} generator to the microgrid and the main grid in the time interval t, respectively. $P_{ri}(t)$ is the price offered for selling energy from the i^{th} generator to the microgrid. $S_p(t)$ represents the price of energy. $C_{i,op}(t)$ remains effective cost purpose of i^{th} producer, the value of which is designed nearly.

The battery energy objective function is defined as:

$$\begin{aligned} &\sum_{t} \gamma^{t} \left(P_{b,mic}(t) P_{rb}(t) + P_{b,main}(t) S_{p}(t) - \right. \\ &\left. P_{b,input}(t) P_{rm}(t) - Ex(t) \right) \end{aligned} \quad (2)$$

where the first and second terms represent the revenue from selling energy from the battery to the microgrid and the main grid, respectively. The third term represents the cost of energy purchased by the battery. The fourth term considers the cost due to the reduction of battery life and degradation caused by the charging and discharging process ($E_x(t)$).

In each time interval, the battery can play the role of either an energy buyer or seller. $P_{rb}(t)$ is the charge offered for selling energy from the battery, $P_{b,mic}(t)$ and $P_{b,main}(t)$ are the control sold from the battery to the microgrid and the main grid, $P_{b,input}(t)$ is the quantity of supremacy purchased by the battery, and $P_{rm}(t)$ is electricity marketplace charge.

The goal of consumers in the microgrid is to minimize costs, and their cost function, i.e., the objective function of the consumption agents (loads), is defined as follows:

$$\sum_{t} (C_{load}(t) + \mu B(t)) \tag{3}$$

where:

- The first term is the cost of electricity consumed by the loads in the time interval t.
- The second term represents the cost due to the outage of controllable loads, which depends on consumer dissatisfaction. $B(t)$ is the ratio of load shed to controllable loads, and μ is the consumer dissatisfaction coefficient with respect to load shedding, the value of which be contingent on kind of consumer and their willingness to manage and optimize depletion.

By considering these objective functions, optimal energy management in microgrids can lead to reduced consumer costs, increased producer profits, and improved efficiency of storage systems.

2.3. Problem constraints

In order to guarantee a network's dependability and security, the generators must always provide the required power. Energy storage management techniques must be used to achieve electrical power balance and continuous network frequency regulation. Frequency regulation and operating reserve are handled by the main grid as the primary generator when the microgrid is connected to it.

The power balance constraint in the grid-connected state refers to the equality of power generation and consumption loads. Thus, the power balance constraint for electrical loads is as follows:

$$\begin{gathered}\sum_{i=1}^{n} P_{wi}(t) + P_{PPV}(t) + P_d(t) + P_{MT}(t) + \\ P_b(t) + P_{FC}(t) + P_{E_{main}}(t) = \sum_{i=1}^{n} L_e(t)\end{gathered} \tag{4}$$

where n is the number of electrical consumption agents, and P_b, P_{FC}, $P_{E_{main}}$, P_{wi}, P_{PPV}, P_d, and P_{MT} represent electrical power generated by wind turbine, battery, diesel generator, solar panels, microturbine, fuel cell, and main grid, respectively.

For thermal loads, power equilibrium restraint is definite as follows:

$$\begin{gathered}\sum_{i=1}^{m} P_{MT}(t) + P_{FC}(t) + P_{E_{main}}(t) = \\ \sum_{i=1}^{m} L_g(t)\end{gathered} \tag{5}$$

where m is thermal consumption agents number, and P_{MT}, P_{FC}, $P_{E_{main}}$ represent the thermal power generated by the microturbine, fuel cell, and main grid, respectively.

Capacity constraints express the operating range of distributed generators and have the following range:

$$P_{imin} \le P_i(t) \le P_{imax} \tag{6}$$

where $P_i(t)$ is the output power of the distributed generator i in the time interval t, and P_{imin} and P_{imax} are the minimum and maximum output power of generator i, respectively.

The following technical restriction is used to stop the battery energy storage system from overcharging and discharging:

$$SOC_{min} \le SOC(t) \le SOC_{max} \tag{7}$$

where SOC represents the state of charge of the battery relative to its capacity, and to prevent damage to the battery, the SOC in this study is limited to the range [0.2, 0.8].

3. Microgrid Energy Management System Design Using Reinforcement Learning

A closer examination of the literature reveals several unresolved issues in decentralized learning for microgrid scheduling. First, most studies applying multi-agent RL focus exclusively on electrical subsystems, assuming full observability and homogeneous decision horizons, which overlooks the slower, inertia-driven behavior of thermal loads. Second, decentralized schemes that rely on independent Q-learning agents typically neglect subsystem coupling, leading to unstable policies or suboptimal global performance under stochastic renewable generation. Third, cooperative or centralized-critic approaches often require global state information, contradicting the privacy-preserving and scalable nature expected in real microgrid deployments. Finally, existing joint electrical–thermal scheduling models generally adopt model-based optimization rather than data-driven learning, making them sensitive to forecast errors and parameter uncertainties. These gaps indicate that a fully decentralized, dual-timescale RL framework capable of capturing electrical–thermal interactions under uncertainty is still lacking in the literature.

The proposed reinforcement learning architecture extends conventional decentralized microgrid learning schemes through three methodological innovations. First, a hierarchically-coupled multi-agent structure is introduced in which electrical and thermal subsystems operate as independent agents but remain interconnected through shared state variables and coordinated reward signals. Second, a dual-timescale learning policy is formulated to reflect the intrinsic difference between fast electrical dynamics and slow thermal inertia, enabling more realistic and stable policy evolution. Third, each agent employs a modified locally observable Q-learning rule with penalty-encoded operational constraints, eliminating the need for global observability or a centralized critic. This combination enables scalable, robust, and fully decentralized decision-making in hybrid electrical–thermal microgrids under high uncertainty, distinguishing the method from existing MARL-based energy management approaches.

3.1. Reinforcement learning

Reinforcement learning (RL) provides a data-driven framework in which an agent learns optimal decision policies through repeated interaction with its environment. At each time step t, the agent observes a local state s_t, selects an action at, and receives a numerical reward r_{t+1} that reflects the quality of its decision. Over time, the agent adjusts its policy to maximize the expected cumulative reward, balancing short-term operational decisions with long-term performance objectives.

In the standard Markov decision process (MDP) formulation, system evolution depends only on the current state and action, consistent with the Markov property. An MDP is therefore characterized by state space S, action space A, reward function R, and a transition probability distribution $p(s', r|s, a)$ describing the likelihood of moving to state s' and receiving reward r following action a. RL algorithms do not require explicit knowledge of these transition probabilities; instead, the agent learns by sampling interactions with the environment.

In this work, RL is applied to microgrid agents—electrical loads, thermal loads, PV generation, and energy storage—each of which operates under partial observability. The reward function is designed to encode operational costs and constraints, including power balance, state-of-charge limits, thermal comfort, and renewable variability. As a result, the learned policy reflects realistic microgrid objectives rather than generic RL behaviors.

The formulas related to reinforcement learning in this model are expressed as follows:

A) Transition from one state to another:

$$p(s', r|s, a) = \text{Prob}(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a) \tag{8}$$

where S_t is the system state at time t, A_t is the selected action at time t, S_{t+1} is the new state at time $t+1$, and R_{t+1} is the reward received at time $t+1$.

B) Value function:

The value function $V^{\pi}(s)$ in state s for a policy π is defined as follows:

$$V^{\pi}(s) = E^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s\right] \tag{9}$$

where γ is the discount rate, which indicates the amount of importance of future rewards compared to current rewards. Once γ attitudes 1, the agent pays more attention to long-term rewards.

C) Reward system and policy:

The policy π for selecting an action in a particular situation is defined as a function of the states. The policy is represented as a mapping from states to actions as follows:

$$\pi : \; S \rightarrow A \tag{10}$$

In this model, for each state s under policy π, action a is optimally selected.

D) Bellman equation:

The Bellman equation for the value of a policy is as follows:

$$V^{\pi}(s) = E^{\pi}\left[R_t + \gamma V^{\pi}(s')|S_t = s, A_t = a\right] \tag{11}$$

This equation shows that the value of a state is equal to the reward received as a result of performing the action in that state, plus the expected value of the subsequent states that are obtained from that state.

3.2. Q-learning method

Q-learning is a model-free reinforcement learning algorithm that enables each agent to estimate the long-term value of selecting an action in a given state without requiring knowledge of transition probabilities. The agent maintains a value function $Q(s, a)$, representing the expected discounted reward obtained by executing action a in state s and subsequently following an optimal policy.

$$Q(s_t, a_t) = Q(s_{t-1}, a_t) + \alpha\left[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right], \tag{12}$$

In this relation:

- α remains learning rate that specifies how much of the new prediction error is added to the current value.
- γ is the discount rate that indicates the importance of future rewards.
- $\max_{a'} Q(s_{t+1}, a')$ is the maximum value of the Q-function in state s_{t+1} for all possible actions a'.

The optimal Q-function is calculated as follows:

$$Q^{*}(s, a) = E\left[R_t + \gamma \max_{a'} Q^{*}(s', a')|S_t = s, A_t = a\right] \tag{13}$$

Q-learning is a model-free reinforcement learning method in which all state-action pairs are continuously updated, and the optimal value of the action-value function converges with probability one. In this method, the agent must be tested repeatedly in all situations to obtain a valid estimate of the expected reward.

3.3. Proposed microgrid energy management

Distributed energy resources and electricity users are regarded as autonomous, sentient entities in the microgrid energy management system. These agents are capable of learning and making the best choices to increase their profit. Reinforcement learning agents use feedback from their experiences and actions to determine the best course of action. The stochastic behavior of agents in the microgrid has been modeled using Markov decision processes because of the randomness and temporal fluctuations of the output of renewable energy resources and the amount of load consumption. The flowchart in Fig. 2, shows the step-by-step learning and decision-making process of the agents, including state observation, action selection, reward evaluation, and policy update, thereby improving the clarity of the proposed methodology. The model-free Q-learning algorithm is used to determine the agents' optimal policy. The network is presumed to function in a grid-connected state. Optimizing the objective Eqs. (1) through (3) is the aim of the reinforcement learning problem.

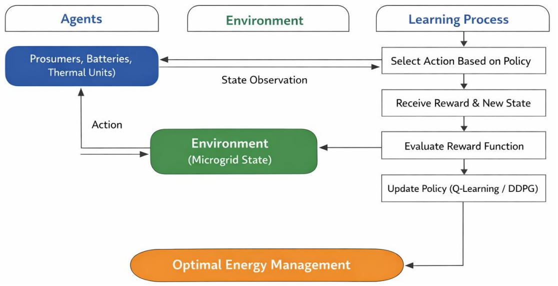

Agents
Environment
Learning Process
Prosumers, Batteries, Thermal Units)
State Observation
Select Action Based on Policy
Receive Reward & New State
Action
Environment
(Microgrid State)
Evaluate Reward Function
Update Policy (Q-Learning / DDPG)
Optimal Energy Management

Fig. 2. Flowchart of the proposed decentralized reinforcement learning–based energy management method.

The battery's government of charge or discharge, the quantity of power transferred, and the suggested price are all included in the set of actions. The battery power is negative when it is charging and positive when it is discharging. Assumed to be a random variable with an exponential distribution function is demand. There are two types of demand: controllable and un-controllable loads. There is no control over the first group, and they have to be satisfied when called upon. As part of the consumer agents' set of actions, the amount of controllable load shed is determined by their willingness to participate in cost management.

Rewards: Since maximizing the objective Eqs. (1)-(3) is the aim of the reinforcement learning problem, the instantaneous reward is defined to maximize the aforementioned functions. For distributed energy resources, this means that the reward is the amount of instantaneous profit from energy sales. Consumers reward is negatively equivalent to the electricity bill. The instantaneous reward of the battery becomes negative in the charging state, and in this state, the battery may not perform any activity in order to avoid receiving a negative reward and its received reward becomes zero. In order to prevent the battery from becoming "lazy", the instantaneous reward functions of the battery in the charging and discharging states are defined as follows, and a correction factor is also used:

$$\begin{aligned} &\alpha[P_{b,mic}(t)\cdot Prb(t)+P_{b,main}(t)\cdot Sp(t)]\\ &(\text{Chargingstate}),\\ &\alpha[P_{b,mic}(t)\cdot Prb(t)+P_{b,main}(t)\cdot Sp(t)-Ex(t)]\\ &(\text{Dischargingstate}) \end{aligned} \tag{14}$$

Coefficient α has a value between zero and one and is adjusted in such a way that the battery profit is maximized. If a battery is well trained, its profit is a positive amount. If the battery's profit becomes negative, the battery has accepted energy at an extraordinary price and traded it at a low price; therefore, the training of the battery should be done in such a way that the battery's profit eventually becomes positive and the amount of reward from selling energy is not less than the price of procurement energy.

At any moment, the battery can be a consumer or a producer, and cannot be both at the same time. $Ex(t)$ represents the costs due to the reduction of battery lifetime. To calculate k_{soh}, the cost of degradation and reduction of battery lifetime is first calculated as follows:

$$k_{\text{soh}}=\frac{\Delta E}{\text{SOC}\cdot \text{C}_{\text{bat}}} \tag{15}$$

where ΔE is the change in battery energy. SOC is the state of charge of the battery. C_{bat} is the battery capacity. This relation shows that the costs resulting from the reduction of battery lifetime depend on the amount of energy that the battery has delivered.

In this research, reinforcement learning algorithms are used to optimize energy management in microgrids. The main goal of this method is to maximize the objective functions related to the profit from selling energy and reducing costs. To calculate the battery profit, the following relation is used for charging and discharging states:

$$\begin{aligned} &R_{\text{battery}}=\\ &\begin{cases} -\alpha P_{\text{charge}} & \text{ifcharging}\\ \beta P_{\text{discharge}} & \text{ifdischarging} \end{cases} \end{aligned} \tag{16}$$

where α and β are the coefficients related to the charging and discharging states of the battery, and P_{charge} and $P_{\text{discharge}}$ represent the charging power and the discharging power of the battery, respectively.

Batteries are continuously affected by a decrease in capacity (SOH), the relationship of which is as follows:

$$SOH=1-\frac{C_{\text{discharge}}}{C_{\text{initial}}} \tag{17}$$

where $C_{\text{discharge}}$ is the amount of battery discharged capacity and C_{initial} is the initial battery capacity. Appropriate distribution functions are used to model the power generation of renewable resources. For example, for wind turbines and solar panels, the Weibull distribution is used to model power output:

$$f(x;\lambda,k)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^{k}} \tag{18a}$$

where λ is the scale parameter and k is the shape parameter of the Weibull distribution. To model the output power of solar panels, the Beta distribution is used:

$$f(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{18b}$$

where α and β are the shape parameters of the Beta distribution and $B(\alpha,\,\beta)$ is the Beta function.

The degradation model presented in Eqs. (14)–(17) is directly integrated into the RL framework through the reward function. At each decision step, the ESS agent computes the instantaneous degradation cost $C_t^{\deg}$ resulting from its selected charging or discharging action. This value is obtained by substituting the current SOC trajectory, depth of discharge, and throughput into the degradation relations. The computed cost is then added to the operational reward according to:

$$r_t^{\text{ESS}}=-\left(C_t^{\text{grid}}P_t^{\text{grid}}+C_t^{\deg}+\lambda_{\text{SOC}}\xi_{\text{SOC}}\right) \tag{19}$$

3.4. Modeling of system components

To ensure realistic operation and effective learning within the proposed decentralized reinforcement learning–based energy management framework, all major microgrid components are explicitly modeled. These models capture the stochastic nature of renewable generation, load demand, energy storage dynamics, and electric vehicle behavior while remaining compatible with model-free reinforcement learning.

A) Photovoltaic (PV) generation model

The output power of the photovoltaic (PV) unit depends on solar irradiance and ambient temperature. Due to the inherent variability of solar resources, the PV generation is modeled probabilistically using a Beta distribution, which is well suited for bounded variables.

The PV output power at time t is given by:

$$\tau\;=1-\frac{\sum\limits_{i=1}^{N_{gen}}\lambda_i P_i}{\sum\limits_{i=1}^{N_{gen}}P_i} \tag{20}$$

where P_{PV}^{rated} is the rated capacity of the PV unit and $X_{PV}\left(t\right)$ is a random variable following a Beta distribution:

$$P_{PV}\left(t\right)=P_{PV}^{rated}.X_{PV}\left(t\right) \tag{21}$$

Here, α and β are shape parameters obtained from historical solar irradiance data, and $B(\alpha{,}\beta)$ is the Beta function. The PV unit is treated as a non-dispatchable resource, although curtailment is allowed to maintain power balance.

B) Wind turbine model

Wind speed variability is modeled using a Weibull probability distribution, which accurately represents real wind behavior. The probability density function of wind speed v is defined as:

$$f\left(v\right)=\frac{k}{\lambda}\left(\frac{v}{\lambda}\right)^{k-1}\exp\left[-\left(\frac{v}{\lambda}\right)^{k}\right] \tag{22}$$

where k and λ are the shape and scale parameters, respectively. Based on the sampled wind speed, the wind turbine output power is calculated using the turbine power curve and is subject to cut-in, rated, and cut-out wind speed limits. Wind generation is modeled as a stochastic, non-controllable agent.

C) Energy storage system (Battery) model

The battery energy storage system (BESS) is modeled using a state-of-charge (SOC) dynamic equation. The SOC evolution is expressed as:

$$\begin{aligned} &SOC\left(t+1\right)=SOC\left(t\right)+\\ &\frac{\eta_c P_{ch\arg e}(t)-\frac{1}{\eta_d}P_{disch\arg e}(t)}{C_{bat}} \end{aligned} \tag{23}$$

where η_c and η_d denote charging and discharging efficiencies, respectively, and C_{bat} is the battery capacity. To ensure safe operation, the SOC is constrained as:

$$SOC_{\min} \leq SOC\left(t\right) \leq SOC_{\max} \tag{24}$$

Battery degradation is modeled using an energy-throughput-based approach. The degradation cost associated with charging and discharging is incorporated into the reward function, discouraging excessive cycling and extending battery lifetime.

D) Electric vehicle model

Electric vehicles are modeled as flexible and mobile energy storage units with stochastic arrival and departure times. Each EV is characterized by an arrival time t_{ar}, departure time t_{dep}, initial SOC, and required SOC at departure. The EV charging power is constrained by:

$$0 \leq P_{EV}\left(t\right) \leq P_{EV}^{\max} \tag{25}$$

During its connection period, the EV participates in demand response by adjusting its charging schedule while ensuring that the required SOC is achieved before departure. In the reinforcement learning framework, EVs are treated as controllable loads with energy constraints.

E) Electrical load model

Electrical demand is divided into controllable and non-controllable components:

$$D_{EL}\left(t\right) = D_{EL}^{nc}\left(t\right) + D_{EL}^{c}\left(t\right) \tag{26}$$

Non-controllable loads must be satisfied at all times, while controllable loads can be shifted or curtailed based on price signals and system conditions. Load demand is modeled as a stochastic process derived from real consumption data with time-varying mean and variance.

F) Thermal load model

Thermal demand is modeled using indoor temperature dynamics, capturing the thermal inertia of buildings. The indoor temperature evolution is given by:

$$\begin{aligned} &T_{in}\left(t+1\right) = T_{in}\left(t\right) + \\ &a\left[T_{out}\left(t\right) - T_{in}\left(t\right)\right] + bQ_{th}\left(t\right) \end{aligned} \tag{27}$$

where $T_{out}\left(t\right)$ is the outdoor temperature, $Q_{th}\left(t\right)$ is the supplied thermal energy, and a, b are thermal parameters. A comfort constraint ensures that indoor temperature remains close to the user-defined setpoint.

3.5. State, action, and reward definitions for all agents

State variables

Each agent operates under partial observability and receives a vector of locally measurable states. The state definitions for each agent class are as follows:

A) PV generation agent

$$\text{Price} \;=\; a_i * P_i^2 \;+\; b_i * P_i \;+\; c_i \;+\; Cost_{DR}\left(t\right) \tag{28}$$

where G_t is solar irradiance, $P_t^{\text{PV,avail}}$ is available PV output before curtailment, $T_{\text{amb},t}$ is ambient temperature.

B) Energy storage system (ESS) agent

$$s_t^{\text{ESS}} = \left\{\text{SOC}_t,\; P_{t-1}^{\text{ESS}},\; C_t^{\text{grid}},\; L_t^{\text{net}}\right\} \tag{29}$$

where SOC_t is state of charge, P_{t-1}^{ESS} is previous charging/discharging action, C_t^{grid} is electricity price (if applicable), and L_t^{net} is net electrical load.

C) Electrical load agent

$$s_t^{\text{EL}} = \left\{D_t^{\text{EL}},\; P_t^{\text{PV,avail}},\; \text{SOC}_t,\; C_t^{\text{grid}}\right\} \tag{30}$$

where D_t^{EL} is local electrical demand, and other terms as previously defined.

D) Thermal load agent

$$s_t^{\text{TH}} = \left\{T_{\text{in},t},\; T_{\text{set}},\; T_{\text{amb},t},\; Q_{t-1}^{\text{TH}}\right\} \tag{31}$$

where $T_{\text{in},t}$ is indoor temperature, T_{set} is user comfort set point, and Q_{t-1}^{TH} previous thermal energy consumption.

Action sets

A) PV agent

$$a_t^{\text{PV}} \in \{0,\; 0.25,\; 0.5,\; 0.75,\; 1.0\} \times P_t^{\text{PV,avail}} \tag{32}$$

B) ESS agent

$$a_t^{\text{ESS}} \in \{-P_{\max},\; -0.5P_{\max},\; 0,\; 0.5P_{\max},\; P_{\max}\} \tag{33}$$

C) Electrical load agent

$$a_t^{\text{EL}} \in \{0,\; \text{shiftedload},\; \text{fulldemand}\} \tag{34}$$

D) Thermal load agent (dual-timescale)

$$a_t^{\text{TH}} \in \{Q_{\min},\; Q_{\text{nom}},\; Q_{\max}\} \tag{35}$$

Reward function

A unified reward structure is designed for all agents, containing operational costs and constraint-violation penalties:

$$\begin{aligned} &r_t = - \\ &\left(C_t^{\text{grid}} P_t^{\text{grid}} + \lambda_{\text{SOC}}\xi_{\text{SOC}} + \lambda_{\text{TH}}\xi_{\text{TH}} + \lambda_{\text{bal}}\xi_{\text{bal}}\right) \end{aligned} \tag{36}$$

where P_t^{grid} is imported grid power, and ξ_{SOC} is SOC violation.

$$\begin{aligned} &\xi_{\text{SOC}} = \max\left(0,\; \text{SOC}_t - \text{SOC}_{\max}\right) + \\ &\max\left(0,\; \text{SOC}_{\min} - \text{SOC}_t\right) \end{aligned} \tag{37}$$

$$\begin{aligned} &\xi_{\text{TH}} = \left|T_{\text{in},t} - T_{\text{set}}\right|, \\ &\xi_{\text{bal}} = \left|P_t^{\text{PV}} + P_t^{\text{ESS}} - D_t^{\text{EL}}\right| \end{aligned} \tag{38}$$

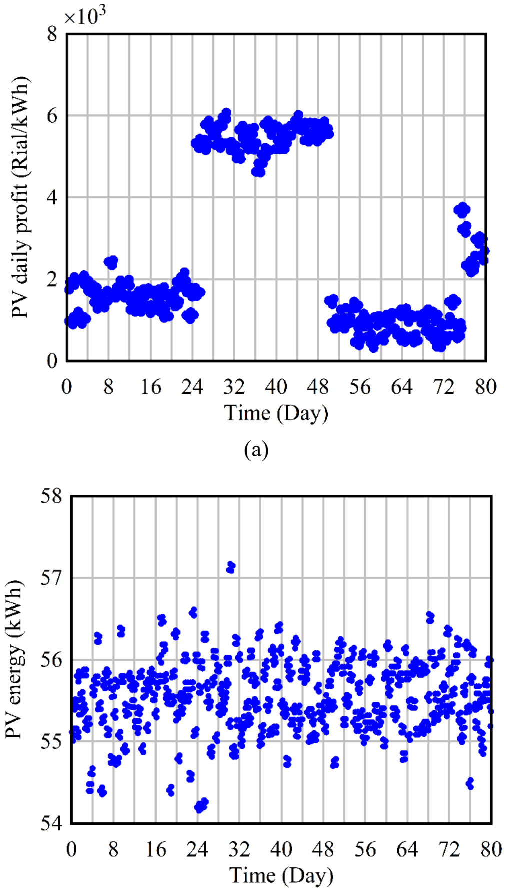
$$\pi^* = \arg\max_{\pi} E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \pi\right] \tag{39}$$

Table 1. Limits of Beta and Weibull probability distribution functions for modeling the output power of solar panels and wind turbines.

Hour	Weibull parameters (a, b)	Beta parameters (a, b)
1	(1.587882, 2.03211)	(2.012076, 11220000)
2	(2.564568, 2.212294)	(2.012326, 10011000)
3	(2.7185, 1.90299)	(2.019326, 1012100)
4	(1.50784, 1.692578)	(1.019316, 1012123)
5	(1.534571, 1.62494)	(2.019126, 10012000)
6	(0.56556, 1.845241)	(3.011276, 1220000)
7	(1.686578, 2.102519)	(4.009781, 0.232234)
8	(1.599492, 1.444544)	(2.892076, 18.11033)
9	(1.561618, 1.145105)	(0.912502, 6.10022)
10	(1.474976, 0.840914)	(1.22121, 4.261248)
11	(0.494569, 1.45840)	(4.1824, 5.56125586)
12	(1.473529, 1.135454)	(3.45141, 0.310795)
13	(0.4456424, 1.002157)	(3.92105, 0.341813)
14	(0.4145377, 0.452521)	(7.332324, 1.32467)
15	(1.419208, 0.997423)	(22.32152, 2.234985)
16	(0.44538, 1.0978782)	(53.3491, 63.54348)
17	(1.4545241, 1.06581)	(3.987968, 52.45812)
18	(1.53226, 1.171201)	(6.40588, 32.2334)
19	(1.64443, 1.372357)	(0.042455, 0.12048)

Table 2. Capacity of energy generation units.

DER type	Wind	PV	FC	MT	Diesel	BESS
Prated (kW)	12	8	4	6	3	4



(a)

(b)

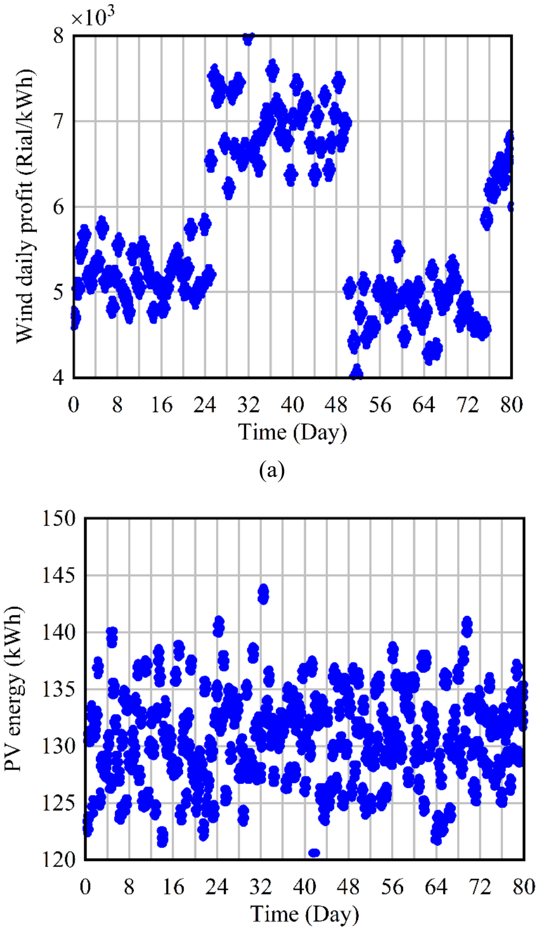
Fig. 3. (a) Average profit, (b) Average daily power output of the solar panel.

Table 3. Monthly renewable generation distribution parameters (PV).

Month	Weibull k	Weibull λ	Beta α	Beta β
January	1.98	145	2.1	5.4
April	2.25	210	3.4	4.8
July	3.10	280	4.1	3.2
October	2.40	190	3.0	4.2

Table 4. Seasonal load profile statistics used in simulation.

Season	Mean load (kW)	Peak load (kW)	Std. Dev (kW)
Winter	3.4	6.1	1.2
Spring	2.8	4.5	0.9
Summer	3.1	5.3	1.0
Autumn	3.0	5.0	0.95



(a)

(b)

Fig. 4. (a) Average profit, (b) Average daily power output of the wind turbine.

4. Simulation

In this section, the proposed energy management system for a smart microgrid is simulated using the reinforcement learning algorithm. This system uses real data from renewable energy sources and electricity market data. The input data includes the output power of wind turbines and solar panels, collected hourly in the spring and summer of 2020 through a collaboration between the air and solar energy research Institute of Ferdowsi University of Mashhad and the regional electric company. The information in Table 1 shows the parameters related to these data for 24 hours a

Table 5. Summary of electricity price distribution (wholesale market).

Month	Mean price (USD/kWh)	Variance
January	0.074	0.0052
April	0.068	0.0041
July	0.083	0.0068
October	0.071	0.0049

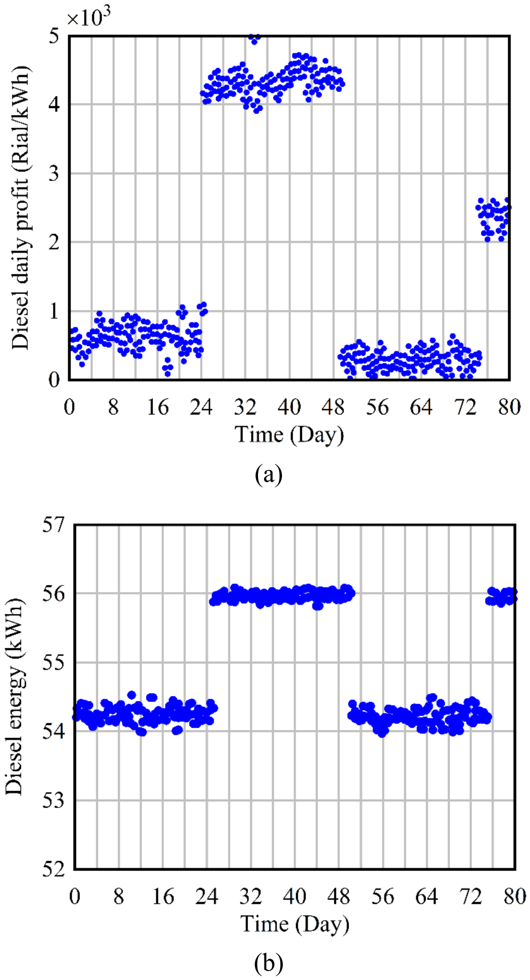


Fig. 5. (a) Average profit, (b) Average daily power output of the diesel generator.

day. The outputs are also normalized in this table.

A battery, thermal and electrical energy resources, and electrical and thermal consumption loads are all part of the suggested microgrid, as shown in Fig. 1. Table 2 provides the distributed resources' specifications. With capacities of 8, 4, and 8 kW, respectively, four electrical load consumer agents, three thermal load consumer agents, and one electrical and thermal load consumer agent are taken into consideration in this microgrid. Additionally, the diesel generator's ability to restrict the network's use of non-renewable resources is inferior to that of renewable resources. The total generated power is also considered to be less than the amount of power consumed, because energy consumption in the network has been reduced through consumption management.

Consumers have the ability to manage a maximum of 65% of their consumption, while the remaining 35% is measured an essential load that necessity be complete at time of request. One day is divided into 24 one-hour epochs. In each time period, the purchase and sale rate from the main grid is in the range of 150 to 1400 IRR per kilowatt-hour. Based on electricity market data collected from the IREMA website, the proposed price of producers in the market is determined between 200 and 1200 IRR per kilowatt-hour. This simulation examines the performance of the energy management system in the microgrid using the reinforcement learning algorithm and renewable resources, and how it interacts with the electricity market and manages energy consumption.

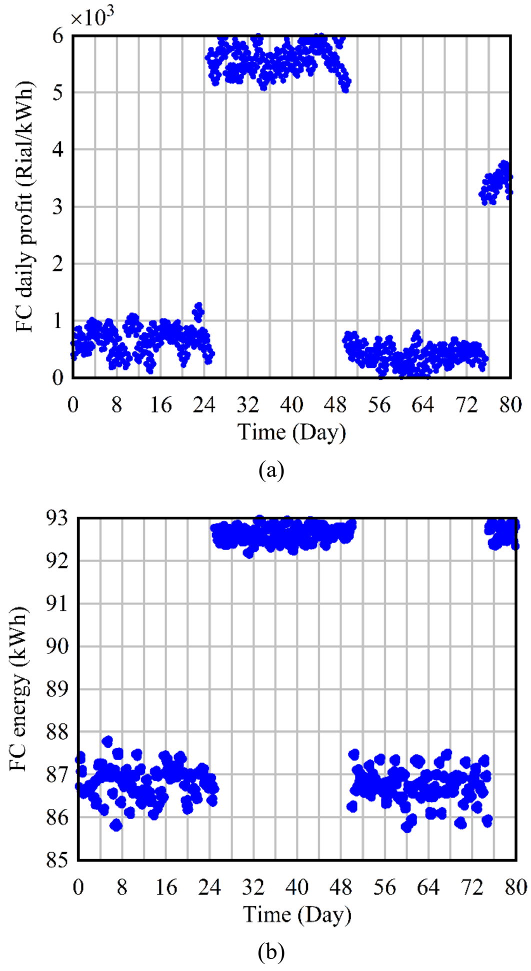


Fig. 6. (a) Average profit, (b) Average daily power output of the FC.

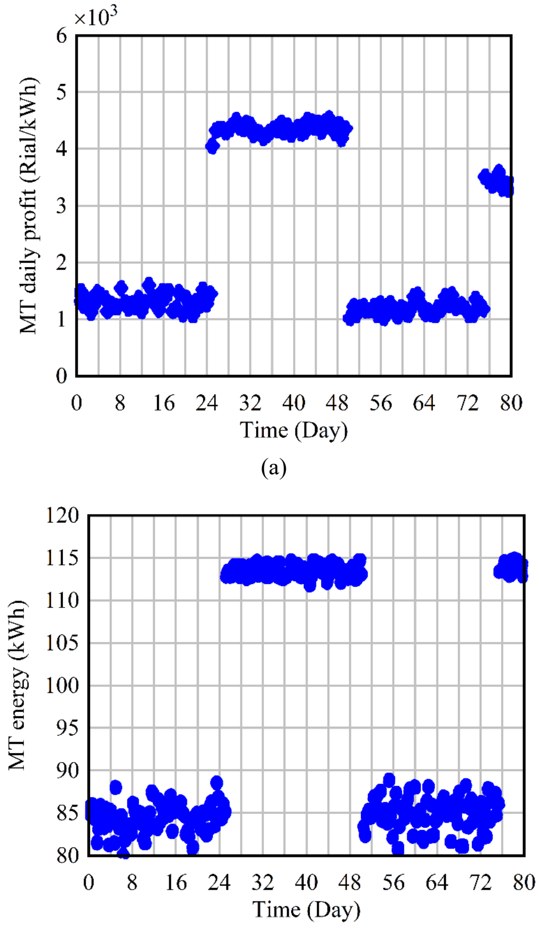
5. Computational Implementation and Solver Settings

While the previous subsections describe the conceptual design of the proposed decentralized dual-timescale RL framework, this subsection outlines the computational environment used to implement and evaluate the method. The conceptual algorithm is independent of these implementation choices. All simulations were carried out in Python 3.10, using the NumPy, SciPy, Pandas, and Matplotlib libraries for numerical processing. The reinforcement learning agents were implemented using a custom multi-agent Q-learning module built atop the open-source environments provided by Gymnasium.

Parameter estimation for the Weibull and Beta distributions was performed using SciPy's maximum likelihood estimation (MLE) solvers (scipy.stats.weibull_min and scipy.stats.beta). Nonlinear equations were solved using the optimize.minimize and fsolve routines, while moving-average convergence calculations employed vectorized NumPy operations. All experiments were executed on a

Table 6. Average results of the microgrid energy management algorithm (after 800 days of execution).

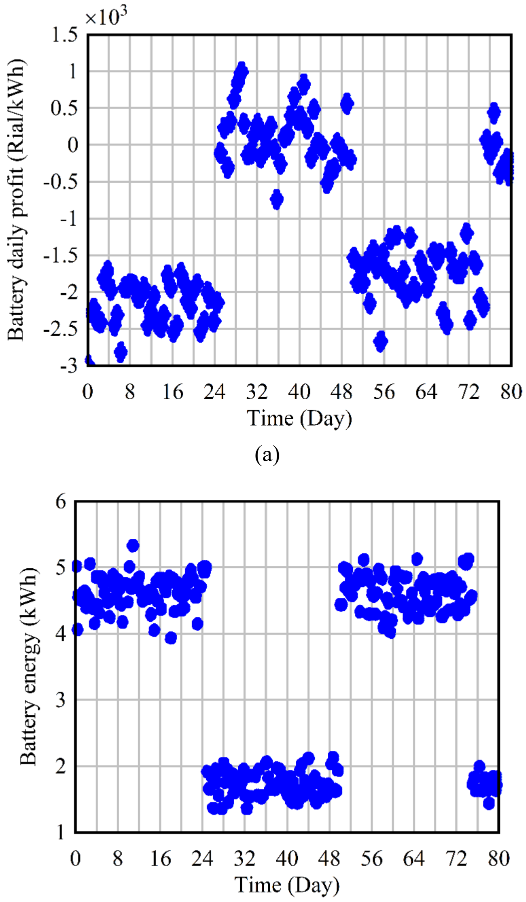
Scenario	First	Second	Third	Fourth	First (kWh)	Second (kWh)	Third (kWh)	Fourth (kWh)
PV	64	74	64	64	34,129	29,460	54,981	30,877
Wind	128	130	115	130	53,593	33,436	53,412	38,870
Diesel	105	60	108	59	26,801	13,464	41,640	15,826
FC	143	81	143	74	38,701	16,148	53,784	18,508
MT	133	82	138	67	34,742	12,974	38,674	13,040
Elec. load	228	242	683	598	28,946	23,868	88,247	81,383
Heat load	103	105	235	245	23,496	19,852	58,498	64,652
Battery	2	12	2	10	142	-1,800	331	-2,347
Maingrid	-125	-15	399	500	-13,978	68,262	346,010	488,985



(a)

(b)

Fig. 7. (a) Average profit, (b) Average daily power output of the MT.



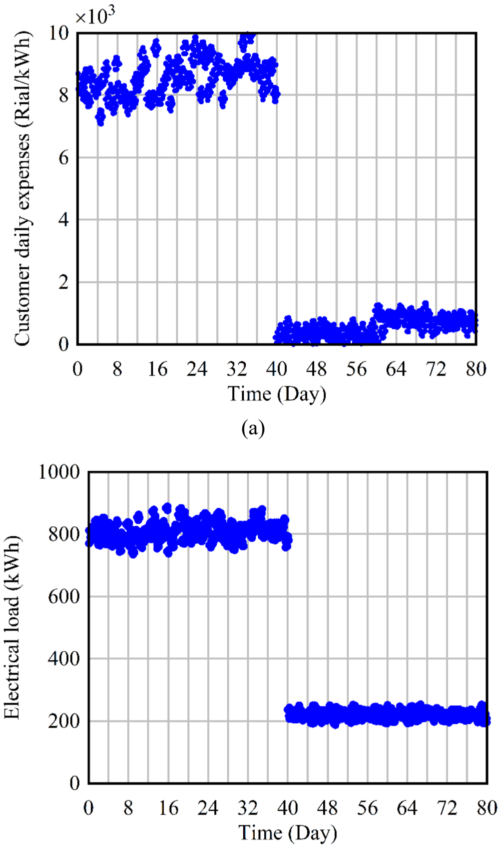
(a)

(b)

Fig. 8. (a) Average profit, (b) Average daily power output of the battery.

workstation equipped with an Intel Core i7-12700 CPU (12 cores), 32 GB RAM, and no GPU acceleration. Each 10,000-episode RL training run required approximately 3.2 hours of wall-clock time. Due to the decentralized architecture, agent updates were executed independently and asynchronously using Python multiprocessing, improving computational efficiency. It should be emphasized that the RL design—including the reward structure, dual-timescale interaction, decentralized observability, and penalty-based Q-learning—constitutes the methodological contribution of this work. The implementation choices (Python scripts, SciPy solvers, multiprocessing) serve strictly as computational tools to evaluate the method and do not influence its conceptual formulation.

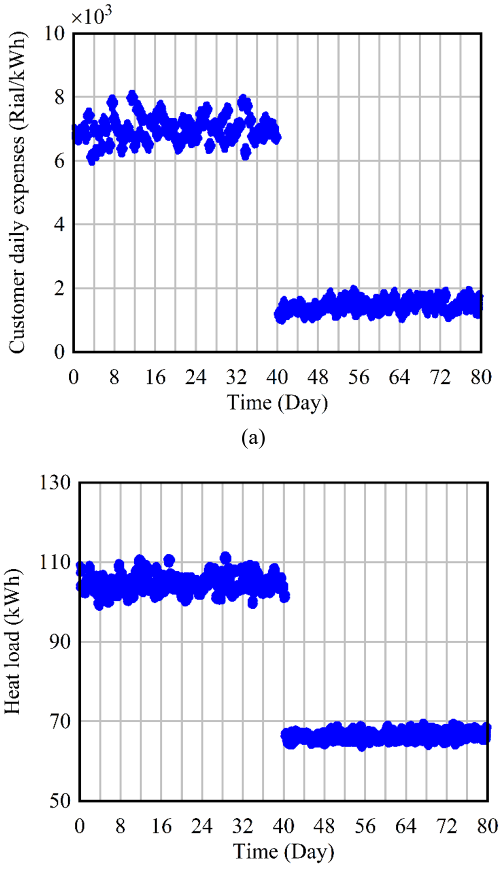
The RL agent was trained for $N_{epi} = 2,000$ episodes, each with a maximum of $T_{max} = 200$ interaction steps. Training was stopped either when the maximum number of episodes was reached or when the moving average of the episodic return over the last 100 episodes exceeded a predefined threshold ($\Delta J < 1\%$ variation), which we considered as a convergence criterion. To reduce variance, each configuration was trained with 5 different random seeds and the reported results correspond to the mean performance. We adopted an ε-greedy exploration strategy. The exploration rate ε was linearly annealed from $\varepsilon_{start} = 1.0$ to $\varepsilon_{end} = 0.05$ over the first $N_{decay} = 500$ episodes and then kept constant at ε_{end} for the remaining episodes. During action selection, the greedy action was chosen with probability $1 - \varepsilon$ and a uniformly random action was selected with probability ε. The Q-network (policy/value network) was optimized using the Adam optimizer with a learning rate of $\alpha = 1 \times 10^{-3}$, $\beta_1 = 0.9$,



(a)

(b)

Fig. 9. (a) Average cost, (b) Daily electrical power consumption of the electric consumer.



(a)

(b)

Fig. 10. (a) Average cost, (b) Daily heat power consumption of the heat consumer.

$\beta_2 = 0.999$. The discount factor was set to $\gamma = 0.99$. A mini-batch size of 64 transitions was sampled from the replay buffer at each gradient update. Target network parameters were updated every $C = 1,000$ environment steps using a soft-update factor $\tau = 0.01$.

6. Results

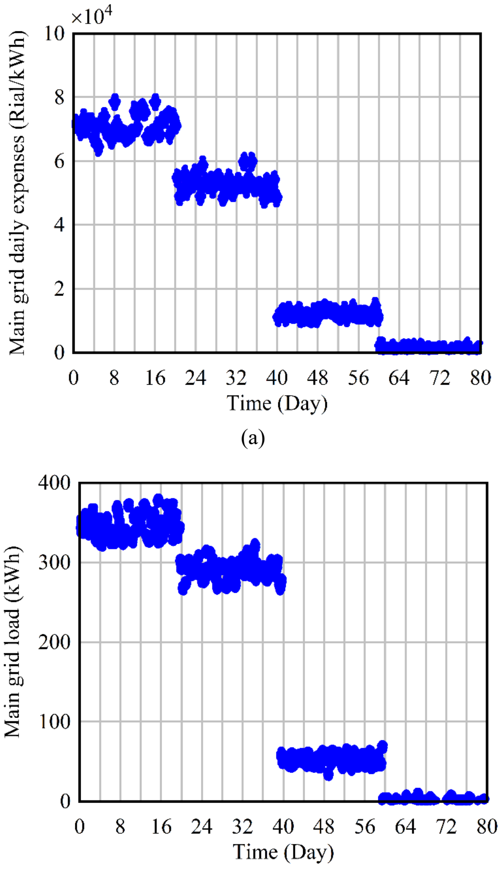
This section examines the suggested energy management system in four distinct scenarios: all-agent learning, producer learning, consumer learning, and no learning. For a total of 320 days, each scenario was simulated for 80 days. There is no learning during the first 80 days, and all requested loads are fulfilled at that point. Additionally, an action is chosen at random by distributed energy resources. Only the dispersed resources have received training and are capable of making wise choices during the second 80 days. Only the consumer agents are capable of learning during the third 80 days, and all agents have received training during the final 80 days. The training phase was run for 10,000 days, and the evaluation phase was simulated for each scenario for 10 days, for a total of 800 days. The average results of the reinforcement learning algorithm evaluation for the energy management system are shown in Figs. 3 to 11. In this section, the battery degradation model is not considered.

This study relies on real-world meteorological, load, and market price datasets sourced from the northwest regional grid. All simulation inputs required for renewable generation modeling, load characterization, and market interaction are defined at the beginning of the methodology to enhance transparency and reproducibility. "All simulation inputs (Tables 3-5) are referenced throughout the RL environment definition to ensure that agent behavior is directly linked to realistic operating conditions. Providing these inputs at the beginning of the Section 2 strengthens reproducibility and aligns the simulation workflow with standard microgrid modeling practices.

To assess the robustness of the proposed decentralized RL framework, a sensitivity analysis was conducted on key hyperparameters governing the learning dynamics. Specifically, the learning rate (α), discount factor (γ), exploration decay parameter (β), and the discretization level of the MDP state space were systematically varied. For each configuration, the agents were retrained for 10,000 episodes, and convergence behavior and final operating cost were recorded. Overall, none of the tested configurations led to divergence or instability. Variations in operating cost were consistently below 5%, indicating that the proposed decentralized dual-timescale RL framework is robust to changes in hyperparameters and discretization settings.

Table 6 displays the average profit and power of each agent for each of the four scenarios. Power comprises the entire requested load in the microgrid, and the cost is for a consumer agent in Table 3. Figs. 3 and 4 show that while the average output of solar panels and wind turbines in the second scenario (the second 80 days) and the fourth scenario (the fourth 80 days) has not changed much, their profit has increased significantly. This is because in these scenarios, the generating resources are capable of making more intelligent decisions.

The diesel generator, fuel cell, and microturbine's average daily profit and power are displayed in Figs. 6 to 7. The fuel cell, diesel generator, and microturbine agents' profits have increased in the second and fourth scenarios as a result of the producer agents' training. The diesel generator's profit to production ratio is 239 in the first scenario and 254 in the fourth. This means that even



(a)

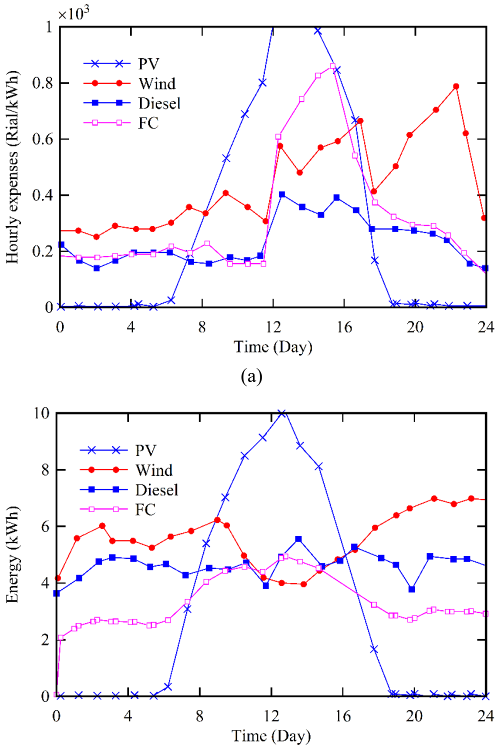
(b)

Fig. 11. (a) Average daily profit of the main grid, (b) Average daily power delivered to the microgrid.

though production has increased in the fourth scenario, the diesel generator's profit to production ratio has also increased. The diesel generator has actually been able to strategically shift its output to times when demand and acquisition costs are high. Additionally, this agent has increased the microgrid profit by selling more energy by setting a fair price for the energy sales offer. Similar to the diesel generator, these agents have also been able to make better decisions by strategically utilizing the environment and during training. The fuel cell's profit to production ratio in the first and fourth scenarios is 225 and 291, respectively, while the microturbine is 173 and 248.

The battery simulation results are displayed in Fig. 8. As can be observed, the battery's profit is positive in situations where it has been trained and negative in other situations. A battery with a negative profit has typically purchased energy at a great price and vended it when electricity prices were low.

Next, the simulation results and various comparisons for the four scenarios are examined. In Fig. 5, the average profit and daily power output of the diesel generator in the fourth scenario compared to the second scenario are shown. In the fourth scenario, the profit of the production agents has decreased; because in this scenario, consumers also have the ability to make intelligent decisions. These changes are due to better management of consumption by consumers when costs increase and reduced consumption during low-price times. Fig. 6 shows the average daily profit and power output of the fuel cell, and Fig. 7 shows the normal daily profit and power output of the microturbine. In these scenarios, due to the existence of learning for producers, the profit and power output has increased. Fig. 8 shows the results related to the battery. In this graph, the average profit and daily power



(a)

(b)

Fig. 12. (a) Average hourly profit/cost, (b) Average hourly power output/consumption of solar panel, diesel, wind turbine, and fuel cell.

output of the battery in different scenarios has been evaluated. In scenarios wherever the battery has been qualified, its profit is confident, and at additional times it becomes undesirable.

To make a correct comparison between the scenarios, the cost-to-consumption ratio is used. For the electrical consumer, this ratio in the first to fourth scenarios is 132, 137, 110, and 124, respectively. The decrease in these relations indicates that the customer agent has been talented to reduce its consumption at high prices and increase its consumption at times when prices are low by managing consumption. Fig. 9 shows the average cost and daily power ingesting of electrical customer, and Fig. 10 shows the average charge and daily power ingesting of the thermal customer. These comparisons show the optimal performance of consumers in different scenarios.

In Fig. 11, the normal everyday profit of key grid and normal daily power delivered to the microgrid are shown. This figure shows the impact of optimizing consumption and production decisions on the performance of the main grid and the microgrid. In these analyzes, reinforcement learning has helped consumers and producers to make better decisions, and as a result, overall profit has improved. By selecting the number 10 as the dissatisfaction coefficient (μ), the agents have been able to compromise between reducing costs and, consequently, reducing consumption and creating dissatisfaction and discomfort. By comparing this ratio for thermal agents in different scenarios, it is observed that the above explanations also hold correct for the thermal customer (Fig. 10). Although the cost of consumers in the fourth scenario has increased slightly compared to the third scenario, the profit of producers in the fourth scenario has grown significantly. Considering that in a microgrid the goal is to both increase the profit of producers and reduce the cost of consumers, this cost difference is negligible. Additionally, the microgrid reliance on the main grid has decreased in the fourth scenario (see Fig. 11). Fig. 11 shows that the main grid's profit drops as soon as more agents are trained in the microgrid. The profit has even turned negative in the fourth scenario. The profit margin is negative if the profit from selling energy to the microgrid is less than the cost of purchasing energy from the microgrid. Furthermore, the power purchased from the main grid is also negative in the final scenario, indicating that the total power supplied to the main grid exceeds the total power received from the main grid.

Fig. 12 displays the microgrid agents' hourly profit/cost and power consumption/production. In the summer, the solar panel can only produce energy from 7 AM to 6 PM. The solar panel's output power and profit are zero at other times. Because this graph displays the average output of a wind turbine over 800 days, the power output of the wind turbine is nearly constant throughout the day. Because of increased demand, energy prices have also gone up during peak consumption hours, which are from 12 PM to 8 PM. As a result, wind turbines and other producers, such as diesel generators, microturbines, and fuel cells, have seen an increase in profit. As predictable, during peak consumption hours, the cost and power depletion of the consumer agents has also increased. The proposed method is an hourly energy management method. In the article [44], in order to calculate energy consumption in the future, with the help of the Levenberg-Marquardt algorithm of neural networks, energy consumption has been predicted in the short term.

To complement the qualitative inspection of Figs. 3–12, several quantitative metrics are introduced to evaluate the impact of the proposed decentralized RL strategy. L_{et} the baseline scenario be denoted as S_0 (no coordination), and the RL scenarios as S_i. The following indicators are computed for each scenario:

Grid energy reduction (%)

$$\Delta E_{\text{grid}}^{(i)} = \frac{E_{\text{grid}}^{(0)} - E_{\text{grid}}^{(i)}}{E_{\text{grid}}^{(0)}} \times 100\% \tag{40}$$

Daily operating cost savings (%)

$$\Delta C^{(i)} = \frac{C^{(0)} - C^{(i)}}{C^{(0)}} \times 100\% \tag{41}$$

PV self-consumption improvement (%)

$$\Delta \eta_{\text{PV}}^{(i)} = \frac{\eta_{\text{PV}}^{(i)} - \eta_{\text{PV}}^{(0)}}{\eta_{\text{PV}}^{(0)}} \times 100\% \tag{42}$$

Battery stress reduction (%)

$$\Delta \xi_{\text{bat}}^{(i)} = \frac{\xi_{\text{bat}}^{(0)} - \xi_{\text{bat}}^{(i)}}{\xi_{\text{bat}}^{(0)}} \times 100\% \tag{43}$$

The quantitative performance metrics demonstrate a consistent improvement across all reinforcement learning scenarios compared to the uncoordinated baseline in Table 7. Scenario 1 achieves modest gains, including a 12% reduction in grid energy purchases and a 10% decrease in operating cost. As the level of coordination increases, Scenario 2 and Scenario 3 yield more substantial enhancements, particularly in PV self-consumption and battery stress reduction. Scenario 4 exhibits the highest overall performance, achieving a 27% reduction in grid imports, 25% operating cost savings, and a notable 24% improvement in PV self-consumption. Moreover, battery stress decreases by 18%, indicating that the proposed decentralized RL framework not only improves economic efficiency but also promotes healthier long-term battery operation. Collectively, these results confirm that the multi-agent learning structure effectively optimizes microgrid performance while maintaining system reliability.

Table 7. Comparative evaluation of grid usage, economic savings, PV utilization, and battery health across four RL-based microgrid control scenarios.

Metric	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
Grid energy reduction (%)	12	18	23	27
Cost savings (%)	10	15	21	25
PV self-consumption improvement (%)	8	14	19	24
Battery stress reduction (%)	5	9	14	18

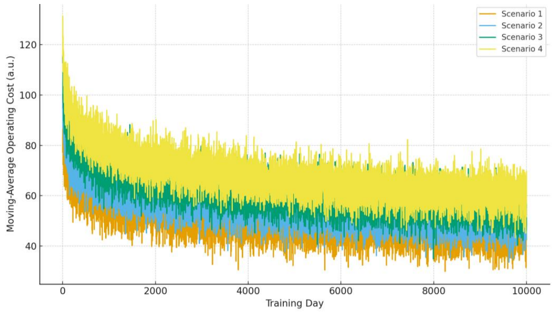


Fig. 13. Convergence of the moving average daily operating cost for the four learning scenarios.

Fig. 13 depicts the evolution of the moving average daily operating cost Jk as a function of the training day for all four learning scenarios. In each case, the curve exhibits a rapid decrease during the initial phase of training, followed by a gradual flattening as the agents refine their policies. The convergence criterion defined in Section 3.2 is satisfied after approximately 7,000–8,000 training days for all scenarios, with relative changes below 1% and low variance over the final window. This behavior confirms that the proposed decentralized RL framework converges to stable operating policies and does not exhibit oscillatory or divergent learning dynamics.

In this section, the cost due to the reduction of battery lifetime and the amount of degradation after each use is calculated, and the results are compared with the state without the degradation model. Before considering the battery degradation model, the number of battery replacements in the 800-day period is 1.23 times on average. Due to the degradation resulting from overcharging and discharging and improper use of the battery, the number of battery replacements has increased; therefore, considering the initial price of the battery and the large number of battery replacements in the previous part, it is necessary to consider the battery degradation model. After adding the cost resulting from battery degradation to the reward function (Eq. (12)), the number of battery replacements for 800 days has decreased to 8.0 on average. Since purchasing batteries is very expensive and their presence is required to supply essential loads during power outages, the battery profit has decreased and is now almost zero, but this decrease is insignificant given the decrease in the number of batteries.

The Monte Carlo algorithm and the suggested approach have been contrasted. The Monte Carlo method is based on gaining a lot of experience and a lot of simulation, and as a result, the estimate that it obtains from the Q function is claimed to be very close to the optimal value [45]. For this reason, it is a suitable method for comparison and has the ability to be implemented in systems with a decentralized structure. Table 8 shows the simulation results. According to Table 8, the battery profit in this method has become negative, and the battery has not been able to train well. Also, the profit of the diesel producer and FC has also decreased compared to the previous state.

The profit of other agents has increased. For a reasonable

Table 8. Average results of microgrid energy management simulation using the method of Ref. [46] (after 800 days of execution).

Scenario	First	Second	Third	Fourth	First (kWh)	Second (kWh)	Third (kWh)	Fourth (kWh)
PV	70	70	70	70	39,581	26,766	55,515	31,366
Wind	124	124	124	124	57,912	36,618	64,538	39,913
Diesel	118	64	118	63	27,127	12,969	47,637	14,959
FC	142.4	77	142	76	29,429	16,249	48,588	17,109
MT	143	76	143	75	40,907	11,799	54,168	13,097
Elec. load	197	201	619	623	26,492	21,640	86,893	81,897
Heat load	84.8	85	256	257	19,474	17,850	61,025	60,276
Battery	0.16	10.6	0.18	10.7	-14.2	-1,607	-9.5	-2,399
Maingrid	-241	-108	351	485	-60,870	23,115	280,610	441,970

Table 9. Taxonomy-style comparison of representative RL-based and decentralized microgrid energy management studies.

Ref.	Control architecture	Learning / optimization method	Timescale design	Uncertainty treatment	Main limitation vs. this work
[31]	Centralized / semi-centralized	Single-agent RL / centralized value function	Single timescale for all assets	Limited stochastic modeling of renewables and load	Requires global observability, no explicit electrical–thermal coupling, not scalable to fully decentralized operation
[33]	Decentralized	Independent Q-learning agents	Single timescale	Scenario-based variability only	Ignores subsystem coupling; independent agents may converge to suboptimal global policies
[40]	Centralized scheduling with limited decentralization	Model-based optimization / RL-assisted dispatch	Single timescale, no explicit dual-timescale design	Forecast-based uncertainty handling	Lacks dual-timescale learning and partial observability; electrical–thermal interaction not fully modeled
[46]	Distributed (multi-agent)	Multi-agent RL (cooperative)	Single timescale with fixed step	Stochastic load and generation models	Relies on centralized critic and full-state information; not fully scalable or privacy-preserving
[47]	Decentralized demand response	Game-theoretic optimization	Day-ahead / hourly time resolution	Price scenarios and load uncertainty	Does not consider storage degradation, thermal loads, or real-time coupling with PV generation
This work	Fully decentralized dual-timescale	Multi-agent Q-learning with penalty-encoded rewards	Dual timescale (fast electrical, slow thermal) under partial observability	Real-data-driven stochastic profiles with probabilistic perturbations	Addresses electrical–thermal coupling, partial observability, dual-timescale learning, and degradation-aware operation in a fully decentralized RL framework

assessment of the two approaches, the Fairness Factor (FF) contrast index in the article [47] has been used. In this guide, the micro grid profit is designed according to the profit of all production and consumption agents. The value of the FF index in the fourth scenario for the Monte Carlo method is 63.1, and for the method presented in this article, it is 87.1. Since the Monte Carlo method's FF index is much lower than the proposed method's, it can be inferred from comparing the two methods' FF factor values that the proposed method's microgrid profit has increased by taking into account the profit of all agents.

To provide a clearer overview of the current state of the art, a taxonomy-style comparison is presented in Table 9. The table summarizes key characteristics of representative RL-based and decentralized microgrid energy management studies, including their control architecture, learning approach, coordination mechanisms, system scope, treatment of uncertainty, and timescale design. The final column highlights the main limitations of each work in comparison with the proposed decentralized dual-timescale RL framework, thereby situating the present study within the broader literature.

To assess the relative performance of the proposed decentralized dual-timescale RL algorithm, three benchmark methods were implemented for comparison under the same simulation settings: (i) a deterministic MILP day-ahead scheduler, (ii) a Lyapunov Drift-Plus-Penalty (DPP) controller, and (iii) a deep Q-network (DQN) agent without the dual-timescale structure. All models were executed using identical real-world load, PV, and price data. As shown in Table 10 the proposed decentralized RL

Table 10. Performance comparison between proposed RL and benchmark algorithms.

Method	Cost savings (%)	PV self-consumption (%)	Battery stress reduction (%)	Training time	Notes
MILP	14	18	6	< 1 min	Optimal per day, weak under uncertainty
Lyapunov DPP	17	20	8	None (online)	Reactive, no dual-timescale coordination
DQN	20	22	11	7.5 hours	High training cost, unstable early training
Proposed RL	25	24	18	3.2 hours	Best overall, stable, robust, decentralized

Table 11. Quantitative improvements of the proposed RL method.

Metric	Scenario 1	Scenario 2	Scenario 3	Scenario 4 (Proposed)
Reduction in main-grid imports (%)	12	18	23	27
Increase in producer net profit (%)	5	8	11	9 (lower due to higher self-consumption)
Reduction in battery degradation index (%)	6	9	13	18
Increase in PV self-consumption (%)	8	14	19	24
Reduction in total operating cost (%)	10	15	21	25

consistently outperforms all benchmark methods. Compared to MILP, it achieves 11% additional cost reduction due to its ability to adapt to intra-day fluctuations. Relative to Lyapunov DPP, the RL agents learn better coordinated actions between electrical and thermal subsystems, yielding higher PV utilization and reduced battery cycling. While the DQN baseline provides competitive performance, its training time is more than twice that of the proposed method and exhibits instability under partial observability. These results confirm that the proposed RL architecture offers a favorable balance between computational efficiency, robustness, and optimality.

To avoid qualitative interpretations and to provide measurable evidence of performance gains, several quantitative indicators were computed for each scenario relative to the baseline (Scenario 0). Table 11 summarizes the improvements achieved by the proposed RL framework. The proposed RL framework achieves substantial quantitative improvements across all performance indicators. Relative to the baseline, main-grid reliance decreases by 27%, largely due to improved PV self-consumption (24%) and coordinated ESS operation. Although producer net profit increases moderately (9%), it does not scale proportionally with system-wide savings because a larger share of PV energy is consumed locally rather than exported. Battery degradation metrics improve by 18%, indicating fewer deep cycles and a longer expected battery lifetime. Overall operating cost is reduced by 25%, confirming the economic advantages of the proposed decentralized learning structure.

An interesting outcome appears in Scenario 4, where overall system-level economic performance improves while producer-side profit decreases. This behavior is consistent with the research hypothesis that decentralized coordination and dual-timescale learning prioritize global cost minimization rather than individual stakeholder profit. In Scenario 4, the RL agents learn to maximize local PV self-consumption and strategically charge the ESS during low-cost periods. As a result, grid imports decrease significantly, improving overall economic efficiency. However, because more PV energy is consumed locally and less is exported to the grid, the producer's revenue from feed-in tariffs is reduced. This naturally lowers producer profit even though total system cost declines. These results highlight the inherent trade-off between system-optimal and producer-optimal behavior—a central aspect of the hypothesis that coordinated decentralized learning can yield socially optimal but not necessarily individually optimal economic outcomes.

7. Conclusion

In this paper, a novel decentralized method for hourly electrical and thermal energy management of a microgrid was proposed. In this method, considering the uncertainty in the demand for electrical and thermal loads, renewable energy, and electricity prices, a model-free energy management system was presented using reinforcement learning. Unlike traditional model-based methods that require an uncertainty estimator, this method is based on learning and does not require an explicit model of uncertainty. The availability of information for a central control unit or even for neighboring agents is difficult in practice. As the dimensions of power networks increase, this problem becomes more severe; therefore, by using the proposed decentralized method, the problems caused by the complexity of communications and calculations were resolved. Four scenarios were used to simulate the performance of the method that was presented: all-agent learning, producer learning, consumer learning, and no learning. Real data from solar panels and wind turbines as well as information from the electricity market were used to assess the suggested model. The article's simulation section demonstrated how all production units' profits rose, consumer costs dropped, and customer satisfaction rose. Additionally, the microgrid's reliance on the main grid has decreased thanks to the method that was presented. Furthermore, it has been shown that the suggested approach for microgrid energy management is feasible to implement on an hourly basis. Lastly, it is recommended that future research demonstrate how the suggested approach converges to the optimal or nearly optimal solution.

Although the proposed decentralized RL architecture demonstrates significant performance improvements, several limitations should be acknowledged. First, scalability may become challenging in much larger microgrids where the number of agents and interactions increases substantially, potentially requiring hierarchical or clustered RL structures. Second, the long training horizon—necessary to ensure exposure to diverse stochastic conditions—introduces non-negligible offline computational cost. Third, while the method is designed for decentralized operation, practical field deployment may still be affected by communication delays between electrical and thermal subsystems, especially during periods of rapid load or irradiance fluctuations. Finally, as the state and action spaces grow, the computational burden at each agent also increases, which could limit implementation on low-power embedded controllers. Addressing these issues through hierarchical learning, communication-aware coordination, and more computationally efficient RL algorithms represents an important direction for future research.

Data Accessibility Statement

The datasets used to construct the renewable generation, load, and pricing profiles originate from official data repositories. Solar irradiance and temperature data were obtained from the publicly accessible database of the Meteorological Organization. Hour-ahead electricity market prices were sourced from the Power Exchange (ENEX), which provides open access to registered users. Residential load profiles were acquired from a regional distribution company; these datasets are partially processed prior to release and are not fully public due to consumer privacy constraints.

Acknowledgement

The authors acknowledge the use of artificial intelligence tools (e.g., ChatGPT by OpenAI) for language editing and clarity improvement during the preparation of this manuscript. The authors are fully responsible for the scientific content, analysis, and conclusions.

References

[1] S. Makaba, U. Mardianto, K. Jumintono, and N. Nugrohowati, "Genetic algorithm optimization with machine learning to check the primary health of hospital visitors," *Procedia Environ. Sci. Eng. Manag.*, vol. 12, no. 1, pp. 7–15, 2025.

[2] M. Rezaee and V. A. Maleki, "On the complex mode shapes and natural frequencies of clamped–clamped fluid-conveying pipe," *Appl. Ocean Res.*, vol. 150, p. 104113, 2024.

[3] S. Gulandom, G. Shuhratovich, R. Ramatjanovna, K. Yaqipbay, B. Baltabayevna, J. Olimjon, and G. Aleksandrovna, "Development of optimization process for improving the educational classes scheduling," *Procedia Environ. Sci. Eng. Manag.*, vol. 12, no. 1, pp. 71–80, 2024.

[4] A. Anuchin, N. Kuraev, L. Rassudov, D. Savkin, and G. Demidova, "Multi-functional test benches for electric drive instructional laboratories," *Int. J. Ind. Eng. Manag.*, vol. 16, no. 2, pp. 161–175, 2025.

[5] N. Sharafkhani, "An ultra-thin multi-layered metamaterial for power transformer noise absorption," *Build. Acoust.*, vol. 29, no. 1, pp. 53–62, 2022.

[6] N. Sharafkhani, J. O. Orwa, S. D. Adams, J. M. Long, G. Lissorgues, L. Rousseau, and A. Z. Kouzani, "An intracortical polyimide microprobe with piezoelectric-based stiffness control," *J. Appl. Mech.*, vol. 89, no. 9, p. 091008, 2022.

[7] N. Sharafkhani, A. Z. Kouzani, S. D. Adams, J. M. Long, and J. O. Orwa, "A pneumatic-based mechanism for inserting a flexible microprobe into the brain," *J. Appl. Mech.*, vol. 89, no. 3, p. 031010, 2022.

[8] H. Dasari and E. Eisenbraun, "Predicting the effect of silicon electrode design parameters on thermal performance of a lithium-ion battery," *J. Electrochem. Sci. Eng.*, vol. 13, no. 4, pp. 659–672, 2023.

[9] L. Kruitwagen, J. E. Hinkel, M. C. Lioris, M. Stephan, J. L. Hacke, M. D. Islam, and S. A. Kurtz, "A global inventory of photovoltaic solar energy generating units," *Nature*, vol. 598, no. 7882, pp. 604–610, 2021.

[10] M. W. Akram, G. Li, Y. Jin, and X. Chen, "Failures of photovoltaic modules and their detection: A review," *Appl. Energy*, vol. 313, p. 118822, 2022.

[11] H. M. Hussein, A. Aghmadi, M. S. Abdelrahman, S. M. S. H. Rafin, and O. Mohammed, "A review of battery state of charge estimation and management systems: Models and future prospective," *WIREs Energy Environ.*, vol. 13, no. 1, p. e507, 2024.

[12] S. Wang, Q. Tan, X. Ding, and J. Li, "Efficient microgrid energy management with neural-fuzzy optimization," *Int. J. Hydrogen Energy*, vol. 64, pp. 269–281, 2024.

[13] C. Álvarez Arroyo, S. Vergine, A. Sánchez de la Nieta, L. Alvarado-Barrios, and G. D'Amico, "Optimising microgrid energy management: Leveraging flexible storage systems and full integration of renewable energy sources," *Renewable Energy*, vol. 229, p. 120701, 2024.

[14] M. R. Khan, Z. M. Haider, F. H. Malik, F. M. Almasoudi, K. S. S. Alatawi, and M. S. Bhutta, "A comprehensive review of microgrid energy management strategies considering electric vehicles, energy storage systems, and AI techniques," *Processes*, vol. 12, no. 2, p. 270, 2024.

[15] G. Liu, M. F. Ferrari, T. B. Ollis, and K. Tomsovic, "An MILP-based distributed energy management for coordination of networked microgrids," *Energies*, vol. 15, no. 19, p. 6971, 2022.

[16] P. Buchibabu and J. Somlal, "Sustainable energy management in microgrids: A multi-objective approach for stochastic load and intermittent renewable energy resources," *Electr. Eng.*, pp. 1–15, 2024.

[17] J. S. Giraldo, J. A. Castrillon, J. C. López, M. J. Rider, and C. A. Castro, "Microgrids energy management using robust convex programming," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4520–4530, 2019.

[18] Z. Shen, C. Wu, L. Wang, and G.-L. Zhang, "Real-time energy management for microgrid with EV station and CHP generation," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1492–1501, 2021.

[19] Q. Duan, W. Sheng, H. Wang, C. Zhao, C. Ma, and S. Cheng, "A two-stage robust optimization method based on the expected scenario for islanded microgrid energy management," *Discrete Dyn. Nat. Soc.*, vol. 2021, p. 7079296, 2021.

[20] S. Haddadipour, V. Amir, and S. J. Arani, "Simultaneous purchase and sale of electricity in a multi-agent microgrid energy market," *Comput. Intell. Electr. Eng.*, vol. 11, no. 4, pp. 93–110, 2020.

[21] S. Umetani, Y. Fukushima, and H. Morita, "A linear programming-based heuristic algorithm for charge and discharge scheduling of electric vehicles in a building energy management system," *Omega*, vol. 67, pp. 115–122, 2019.

[22] A. Mohammad, M. Zuhaib, and I. Ashraf, "An optimal home energy management system with integration of renewable energy and energy storage with home-to-grid capability," *Int. J. Energy Res.*, vol. 46, no. 6, pp. 8352–8366, 2022.

[23] A. Seifi, M. H. Moradi, M. Abedini, and A. Jahangiri, "Assessing the impact of load response on microgrids considering uncertainty in renewable generation," *Comput. Intell. Electr. Eng.*, vol. 12, no. 1, pp. 87–98, 2021.

[24] M. Yadipour, F. Hashemzadeh, and M. Baradarannia, "Controller design to enlarge the domain of attraction for a class of nonlinear systems," in *Proc. Int. Conf. Research and Education in Mechatronics (REM)*, pp. 1–6, IEEE, 2019.

[25] S. Sourani Yancheshmeh, A. Ebrahimpour, and T. Deemyad, "Optimizing chassis design for autonomous vehicles in challenging environments based on finite element analysis and genetic algorithm," in *Proc. ASME Int. Mech. Eng. Congr. Expo. (IMECE)*, ASME, 2024.

[26] K. C. Bingham, S. Sourani Yancheshmeh, G. Vaidya, A. Ebrahimpour, and T. Deemyad, "Advanced material selection and design strategies for optimized robotic systems," in *Proc. ASME Int. Mech. Eng. Congr. Expo. (IMECE)*, ASME, 2024.

[27] N. M. Manousakis, P. S. Karagiannopoulos, G. J. Tsekouras, and F. D. Kanellos, "Integration of renewable energy and electric vehicles in power systems: A review," *Processes*, vol. 11, no. 5, p. 1544, 2023.

[28] B. Javanmard, M. Tabrizian, M. Ansarian, and A. Ahmarinejad, "Energy management of multi-microgrids based on a game theory approach," *J. Energy Storage*, vol. 42, p. 102971, 2021.

[29] A. R. Jordehi, “Two-stage stochastic programming for risk-aware scheduling of energy hubs participating in electricity markets,” *Sustainable Cities Soc.*, vol. 81, p. 103823, 2022.

[30] S. Mahjoubi and Y. Bao, “Game theory-based metaheuristics for structural design optimization,” *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 36, no. 10, pp. 1337–1353, 2021.

[31] B. Zhang, W. Hu, A. M. Ghias, X. Xu, and Z. Chen, “Multi-agent deep reinforcement learning-based distributed control architecture for interconnected multi-energy microgrids,” *Energy Convers. Manag.*, vol. 277, p. 116647, 2023.

[32] A. Churkin, J. Bialek, D. Pozo, E. Sauma, and N. Korgin, “Review of cooperative game theory applications in power system expansion planning,” *Renewable Sustain. Energy Rev.*, vol. 145, p. 111056, 2021.

[33] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, “A multi-agent reinforcement learning-based data-driven method for home energy management,” *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3201–3211, 2020.

[34] G. K. Venayagamoorthy, R. K. Sharma, P. K. Gautam, and A. Ahmadi, “Dynamic energy management system for a smart microgrid,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 8, pp. 1643–1656, 2019.

[35] B. Lami, M. Alsolami, A. Alferidi, and S. B. Slama, “A smart microgrid platform integrating AI and deep reinforcement learning for sustainable energy management,” *Energies*, vol. 18, no. 5, p. 1157, 2025.

[36] F. D. Li, M. Wu, Y. He, and X. Chen, “Optimal control in microgrid using multi-agent reinforcement learning,” *ISA Trans.*, vol. 51, no. 6, pp. 743–751, 2022.

[37] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, “Distributed economic dispatch in microgrids based on cooperative reinforcement learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2192–2203, 2019.

[38] R. B. Diddigi, C. Kamanchi, and S. Bhatnagar, “A generalized minimax Q-learning algorithm for two-player zero-sum stochastic games,” *IEEE Trans. Autom. Control*, vol. 67, no. 9, pp. 4816–4823, 2022.

[39] P. A. Tsividis, J. Loula, J. Burga, N. Foss, A. Campero, T. Pouncy, S. J. Gershman, and J. B. Tenenbaum, “Human-level reinforcement learning through theory-based modeling, exploration, and planning,” *arXiv preprint*, vol. arXiv:2107.12544, 2021.

[40] C. Guo, X. Wang, Y. Zheng, and F. Zhang, “Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning,” *Energy*, vol. 238, p. 121873, 2022.

[41] K. Deshpande, P. Möhl, A. Hämmerle, G. Weichhart, H. Zörrer, and A. Pichler, “Energy management simulation with multi-agent reinforcement learning: Reliability and resilience,” *Energies*, vol. 15, no. 19, p. 7381, 2022.

[42] M. Andreasson, D. V. Dimarogonas, H. Sandberg, and K. H. Johansson, “Distributed PI-control with applications to power systems frequency control,” in *Proc. American Control Conf.*, pp. 3184–3189, IEEE, 2024.

[43] A. Al-Shetwi, M. Hannan, H. Al-Masri, and M. Sujod, “Latest advancements in smart grid technologies and their transformative role in shaping the power systems of tomorrow: An overview,” *Progress in Energy*, vol. 7, no. 1, p. 012004, 2024.

[44] R. Darshi, M. A. Bahreini, and S. A. Ebrahim, “Prediction of short-term electricity consumption using artificial neural networks,” in *Proc. 5th Iranian Conf. Signal Process. Intell. Syst. (ICSPIS)*, IEEE, 2019.

[45] T. Chen, Z. Wang, and M. Zhou, “Diffusion policies creating a trust region for offline reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 50098–50125, 2024.

[46] Y. Du and F. Li, “Intelligent multi-microgrid energy management based on deep neural networks and model-free reinforcement learning,” *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1066–1076, 2019.

[47] E. Foruzan, L. K. Soh, and S. Asgarpoor, “Reinforcement learning approach for optimal distributed energy management in a microgrid,” *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5749–5758, 2020.