Contents lists available at ScienceDirect

ISA Transactions

journal homepage: www.elsevier.com/locate/isatrans

Research article

A novel deep reinforcement learning based business model arrangement for Korean net-zero residential micro-grid considering whole stakeholders' interests

Lilia Tightiz, Joon Yoo [*]

School of Computing, Gachon University, 1342 Seongnam-daero, Sujeong-gu, Seongnam-si, 13120, Gyeonggi-do, Republic of Korea

ARTICLE INFO

Article history:
Received 26 May 2022
Received in revised form 28 November 2022
Accepted 10 December 2022
Available online 13 December 2022

Keywords:
Business model
Deep reinforcement learning
Demand response
Energy management system
Mixed integer nonlinear programming
Net-zero building

ABSTRACT

In this paper, we put forward a deep reinforcement learning (DRL) based energy management system (EMS) solution for a typical Korean net-zero residential micro-grid (NZR-MG). We model NZR-MG EMS to extract a profitable business model that respects whole stakeholders' interests and meets Korean power system regulations and specifications. We deployed the value-based DRL technique, dual deep Q-learning (DDQN), as a solution for our EMS problem since of its simplicity, stability in the learning process, and non-dependency on hyper-parameter selection compared to actor–critic methods. Due to the implementation of mixed-integer nonlinear programming (MINLP) to solve the reward function in this paper, DDQN, despite other DRL methods, provides precise, explicit, and meaningful rewards. In addition to encouraging the agent to choose profitable actions, this approach releases the proposed DRL-based method from the hindrance of redesigning the reward function experimentally in any future extension of the environment elements. Moreover, attaching transfer learning (TL) to the process of training DDQN agent defeat the MINLP imposed latency in training convergence. An extensive benchmark is proposed to test the superiority of the proposed method versus other DRL algorithms.

1. Introduction

Over the past decade, there has been a massive movement worldwide to use renewable energy sources (RESs) under the supervision of micro-grids to control their random behavior [1]. The utilization of RESs in the micro-grid is attracting considerable attention to study different aspects of micro-grid network performances, including electrical parameters control to provide stability and EMS optimization. Micro-grid EMS plans to address objectives such as reducing cost of power generation, consumption, and maintenance, participating in demand response programs, etc. [2]. Despite the importance of stakeholder satisfaction in energy efficiency planning of the micro-grid, there remains a paucity of evidence on providing the EMS based on the precise business model [3]. Business models for micro-grid deployment in each country need to be customized to address dominant types of micro-grid penetration and its power system structure and regulations.

Korea increasingly has substituted fossil fuels with RESs after the ministry of trade, industry, and energy announced incentive policies for developing RESs-related industries in 2016, along with the liberalization of the electricity generation market [4]. In Korea, new RES deployment strategies began with the micro-grid-based green islands introduction and evolved via the construction of NZR-MG in several cities [5]. Among the globally accepted business models of micro-grid owner financing and pay-as-you-go are the most accepted approach. The facility owner constructs the micro-grid, funds the project, then manages the assets in the owner financing & maintenance model. Grid-connected campus micro-grids follow this business model. On the other hand, pay-as-you-go is sponsored by world-class financial organizations and used in remote locations, with customers paying for their energy consumption [6]. Given the definitions and objectives of each model, the combination of both mentioned models can meet the Korean style micro-grid penetration in the metropolitan areas. After the micro-grid construction, private sectors maintain, control, and improve micro-grid infrastructure according to the owner financing & maintenance model. In this structure, consumers only pay for their energy consumption according to the pay-as-you-go model. This combined business model will generate fresh insight into micro-grid's EMS by considering benefits for whole stakeholders and enhancing the micro-grid structure to access maximum profit.

On the other hand, stochastic and model-free characteristics of the micro-grid environment call for an optimization model that supports sequential and online learning. As a subset of machine learning, reinforcement learning (RL) involves an agent that

[*] Corresponding author.
E-mail address: joon.yoo@gachon.ac.kr (J. Yoo).

https://doi.org/10.1016/j.isatra.2022.12.008

learns from observing the outcomes of its actions in a given environment. The most significant recent advance in the RL field is the model-free approach that obtains the optimal solution without getting access to the precise model of the environment. RL is also combined with the deep neural network (DNN) to solve high-dimensional problems and is called DRL. The other benefit of deploying DNN for micro-grid is providing the pattern of environment elements that behave stochastically. In this way, DRL will release from the necessity of forecasting those parameters' future status. Therefore, in this paper, we will hire DRL to schedule EMS for NZR-MG.

2. Literature review

Since this paper's objective is to provide a DRL-based solution that supports profitable EMS scheduling for whole stakeholders, we categorized related studies into three different approaches. Firstly, studies that solved the EMS problem using standard DRL methods since they support high-dimensional, model-free, and stochastic characteristics of the micro-grid environment [7–9]. Secondly, several researchers have combined DRL methods or used model-based optimization to enhance the multi-dimensional support characteristic of DRL and accelerate convergence optimization [10–13]. Finally, a number of studies have focused on profit-driven optimization of micro-grid EMS without applying DRL techniques. However, all of the aforementioned efforts arrange profitable EMS according to the general insight or tendentious one of micro-grid stakeholders [14–16].

With the aim of an online solution for residential micro-grid elements, including electric vehicles (EVs), smart appliances, and photovoltaics (PVs), a deterministic policy gradient (DPG) was hired in [7]. The primary objectives of this study were to reduce peak demand and minimize power demand costs. Bui et al. in [8] deployed DDQN to schedule the energy storage systems (ESSs) community of micro-grids to satisfy consumers in islanded mode by preventing the shed of critical loads and minimizing the cost of power generation in grid-connected mode. Tightiz et al. in [9] compared soft-actor critic (SAC) and deep deterministic policy gradient (DDPG) techniques performances for micro-grid EMS. Authors in this paper defined an innovative cost function to maximize profit for micro-grid owners while this cost function considers utility grid interests giving priority to supporting the power system in contingency situations with micro-grid higher reliable energy resources. However, despite the priority of DRL in providing a model-free solution, there is a dedicated weak point in RL methods, which is the sensitivity of speed and accuracy of the system learning convergence to the definition of the proper reward function. Therefore, the reward function was defined experimentally with trial and error to converge the learning process to sensible results.

Nakabi et al. in [10] compared a diverse set of DRL methods executions such as deep Q-network (DQN), asynchronous advantages actor–critic (A3C), proximal policy optimization (PPO), etc. in the residential micro-grid optimization participating in demand response. Authors in this paper deployed replay buffer memory of previous actions to control correlation in action selection for A3C and PPO, which improves convergence to the optimal policy for both algorithms. Long short term memory neural networks (LSTM) were hired in [11] to determine patterns of PV, wind turbine (WT), and load from historical data of a residential micro-grid. The evoked patterns arranged an environment for a model-based RL to minimize the cost of power generation and consumption. This paper also considered the load power flow limitations of the proposed electricity network. Yoldas et al. in [12] considered daily and emission costs minimization of a campus micro-grid with Q-learning. Using MINLP to set up an EMS for the micro-grid, the authors considered PV, ESS, and diesel generator (DG) constraints to minimize daily and emission costs for the micro-grid. Micro-grid operators planned micro-grid EMS without observing EV and air conditioner data in [13] by utilizing vectorized advantage actor–critic (A2C) to respect the consumers' privacy. The authors, in this paper, to provide a stable learning process deployed the gated recurrent unit (GRU) neural network method to estimate the value function of actions and policy estimation of critic and actor networks. Deploying GRU, which uses fewer gates, resulted in lower memory consumption and faster computation than LSTM. Although there are signs of progress in speed convergence improvement of DRL methods learning process in [10,11,13], implicit reward function definition remains a challenge.

Ref. [14] is one of the initial studies on evoking profit-driven residential micro-grid scheduling that concerned various stakeholders' interests in their solution. In this proposed business model, the micro-grid traded power with the utility grid through the supervisory of the aggregator based on a set of rules. Furthermore, smart home appliances contribute to demand response for peak-shaving, respecting the utility grid profit. This study identified the optimal number of housing units, size of battery energy storage systems (BESS), and area of PVs by utilizing space exploration techniques. However, this study concentrated on the design alternatives before the micro-grid construction using space exploration methodology. In addition, the effect of EVs in residential micro-grid optimal scheduling is not negligible these days compared to the time of publication of the paper [17]. A micro-grid supplier in [15], equipped with the dispatchable natural-gas-based power generator, guarantees the reliability of customer power supply by offering backup power to the micro-grid customer in the utility grid's absence. In this paper, when the power generation cost, including fuel, emission, and maintenance costs, are lower than selling power to the utility grid, the supplier was scheduled to sell power in the electricity market. Monte Carlo simulation evaluated this commercial arrangement of micro-grid by utilizing Texas power system reliability indexes and the power market. Qu et al. in [16] explored a community-scale micro-grid business model for an industrial park. The authors, in this paper, without specifying the optimization method, introduced a platform for the green micro-grid project planning and operation. This study scheduled BESSs and demand response for maximum profit considering the lowest emissions and electricity charges in different scenarios for PV performances, including the highest, the lowest, and fluctuation in power generation and micro-grid operation in islanded and grid-connected modes. Although [14–16] considered the micro-grids business model in EMS cost function arrangement, there have been few empirical investigations on different stakeholders' interests according to the power system structure and regulations of the understudy region of the micro-grid.

In this paper, we deploy DDQN as a DRL method to provide a model-free and online scheduling that support stochastic and uncertain characteristics of the RES-based micro-grid and release from forecasting tools deployment. DDQN is a value-based DRL technique that tackles DQN's learning process instability. DDQN utilizes two separate networks for Q-value estimation to eliminate the correlation between the target value and the estimated value. However, DDQN supports discrete action space problems. There is also the enhanced family of the DRL, which is an actor–critic that supports continuous action space. This specification is suited to BESS micro-grid element performance. However, actor–critic techniques such as DDPG and A3C utilized in [10,13] are subjected to unstable and unreliable learning processes due to the wide range of hyper-parameter calibration requirements [18]. Hence, to make a tradeoff between resolution, simplicity, and accessibility in the learning process, we hired DDQN. Furthermore, in DRL, learning is not usually accurate if reward functions are based on actual cost functions. Standard DRL methods do not produce a precise learning process using realistic reward definitions and require trial-and-error to develop the implicit reward that meets the action selection objectives. In the case of profitable micro-grid EMS unit arrangement, an MINLP approach is a proven efficient technique to model actual cost function. The complexity of solving the MINLP model highlights the problematic issue of realistic reward determination for the DRL arrangement of the micro-grid. Therefore we develop a reward estimator to hire explicit rewards by attaching the MINLP solver to the DDQN. This approach will accelerate policy optimization and, consequently learning process. Additionally, it is feasible to hire the proposed method for any future system extension with minimal effort due to this realistic reward definition compared to hiring implicit rewards. Another disadvantage of hiring continuous action space-based methods such as DDPG is their combination with the MINLP solver will result in massive action space. The discrete action characteristic of the DDQN algorithm supports the strategy-based solution that we set by deploying MINLP. This approach appears to improve the accuracy of the system performance comparing other methods. However, the convergence speed will be subject to latency by combining DDQN with model-based solutions. To conquer this hindrance, we deployed TL to control the extreme behavior of the system, such as state of the charge (SoC) limitation and power balance. Our approach involves taking learned parameters of different EMS task levels to our target task, which accelerates the learning process and accuracy of DDQN action selection.

Furthermore, this study presented here is one of the first investigations to focus specifically on EMS arrangement for the NZR-MG business model based on Korean power system structure and regulations. This research fills the gap in the literature on how it is possible to enhance existing NZR-MG EMS to be profit-driven by respecting the whole stakeholder's benefits. Therefore, we consider an NZR-MG with the same load pattern as a residential complex in metropolitan areas in Korea. We stipulate the baseline model of this NZR-MG deploys commonly distributed energy resources (DER) in Korea's net-zero buildings, including PVs and fuel cells. We develop the baseline model by considering EV stations and consumers' participation in demand response to study their effect on making the micro-grid business model profitable for whole stakeholders and explore new opportunities to accelerate increase micro-grid and holistically widen the horizon of NZR-MG utilization in Korea. Hence, the contributions of this paper to providing EMS that fits Korea's power system structure and regulations are as follows.

- Business model arrangement for NZR-MG considering whole stakeholders profit;
- Development of the NZR-MG model with demand response and EV contribution;
- Implementation of profit-driven EMS for NZR-MG through online learning by deploying DDQN;
- Arrange a novel DRL solution for our NZR-MG model based on DDQN attaching MINLP as a reward estimator and TL for accelerating the learning process.

The remainder of this paper will be as follows: Section 3 provides a business model for NZR-MG proportional to the Korean power system structure and regulations. Section 4 investigates elements constraints and modeling of the understudy NZR-MG. We devote Section 5 to the development procedure of the hired DRL technique in response to the proposed business model objectives. In Section 6, we examine the efficiency of the proposed technique in different cases of NZR-MG by comparing its performance with benchmark solutions. Ultimately, this paper ceases in Section 7 as a conclusion.

3. NZR-MG business model arrangement

The initial objective of the micro-grids advent was to avoid power system generation and transmission lines extension by deploying DERs to electrify remote places. Integrating RESs into micro-grid enables this solution to address environmental concerns. In the Korean peninsula, initially, islanded micro-grid was attractive to provide green islands, and later grid-connected micro-grid were deployed through the cities by implementing NZR-MGs. The NZR-MG projects in Korea often involve a group of nearby residential buildings. Net-zero buildings equipped with intelligent energy management solutions and generating heating from renewable energy as objectives of the Korea 2050 net-zero plan accelerated by improving NZR-MGs from self-sufficient to electricity market participants [19]. Therefore, NZR-MG will offer profit to their stakeholders besides reducing electricity production costs. Since, in this paper, we considered the implemented NZR-MG project in Korea, we will define a business model that improves this residential micro-grid to provide profit. The micro-grid business model includes planning, deployment, and operation of micro-grids to meet its tactical goals. Ref. [20] offers a variety of business models for micro-grids and their components [20]. We adapted the business model framework for micro-grid in [20] to our understudy NZR-MG and planned the business model framework as shown in Fig. 1. Several elements that are determinants in the NZR-MG business model, including NZR-MG (R), technology provider (T), utility grid service provider (U), and micro-grid service provider (S). As we design a profit-driven micro-grid business model, it is required to analyze the interactions between elements and determine the associated costs and revenue. As seen in Fig. 1, let $\mathbf{G} = (V, E)$ denotes the graph of our NZR-MG business model where V is the set of business model elements and E is the set of relationships between elements based on costs and income. Given the definition of $\mathbf{G}$, we can calculate the profit of our business model arrangement at time t as follows.

$$\mathcal{P}(t) = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathit{Profit}(K_a^i|\zeta_{i,j})(t), \tag{1}$$

$$\mathit{Profit}(K_a^i|\zeta_{i,j})(t) = \sum_{i=1}^{N}\sum_{j=1}^{N}(\mathit{Revenue}(K_a^i|\zeta_{i,j})(t) - \mathit{Cost}(K_a^i|\zeta_{i,j})(t)). \tag{2}$$

where K_a is the key action of the proposed business model elements, N is the number of NZR-MG elements, and $\mathit{Profit}(K_a^i|\zeta_{i,j})$ is the amount of earning from each activity after distracting the activity cost.

$\mathit{Profit}(K_a^i|\zeta_{i,j})$ depends on the actions and relationships between elements of our business model. We assign weights to the relationship between entities and recognize adjacent matrices for the graph $\mathbf{G}$, which is,

$$\mathbf{M} = \begin{array}{c|cccc} & R & I & T & S \\ \hline R & 0 & 0 & \zeta_{13} & \zeta_{14} \\ I & 0 & 0 & \zeta_{43} & \zeta_{24} \\ T & \zeta_{31} & \zeta_{34} & 0 & \zeta_{32} \\ S & \zeta_{41} & \zeta_{42} & \zeta_{23} & 0 \end{array}$$

where,

$\zeta_{13} = \{Cost^{-}_{deg}\}$,
$\zeta_{14} = \{D^{-}, D^{+}\}$,
$\zeta_{23} = \{Investment\ on\ Infrastructure^{+}\}$,
$\zeta_{24} = \{\pi^{+}_{DR}, \pi^{-}_{Purchasing}, \pi^{+}_{Selling}\}$,
$\zeta_{31} = \{Cost^{-}_{Ownership}\}$,
$\zeta_{32} = \{Cost^{-}_{Infrastructure}\}$,

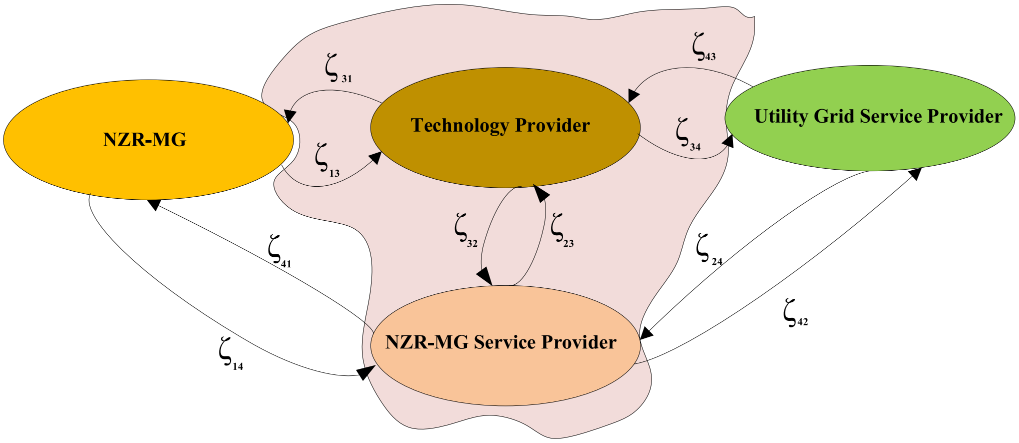


Fig. 1. The business model framework for NZR-MG.

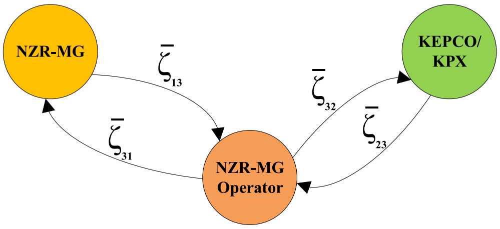


Fig. 2. The business model framework for NZR-MG with subtracting technology mapping.

$\zeta_{34} = \{\pi^{+}_{env}\}$,
$\zeta_{41} = \{Bill\ reduction^{+}\}$,
$\zeta_{42} = \{Peak\ shaving^{+}\}$,
$\zeta_{43} = \{Emission\ reduction^{+}\}$.
$Cost^{-}_{deg}$ is the cost of utilized technology degradation, D^{-} is the customer demand, D^{+} is the customer demand reduction from utilizing technology, π^{+}_{env} is the policy of the institution structure supervisory level to encourage technology contribution in net-zero planning, π^{+}_{DR} is incentives for taking part in demand response, $\pi^{-}_{Purchasing}$ and $\pi^{+}_{Selling}$ are policies for trading power with the utility grid. However, other relationships such as $Cost^{-}_{Ownership}$, $Cost^{-}_{Infrastructure}$, $Bill\ reduction^{+}$, $Investment\ on\ Infrastructure^{+}$, $Peak\ shaving^{+}$, and $Emission\ reduction^{+}$ are clear. It is noted that the sign + shows the revenue while - shows the cost.

We coarsen the graph by merging technology with the third party and arranging a super vertex. We relaxed ζ_{34} and ζ_{43} on trading power policies. Since the operator is responsible for maintenance and investment in the micro-grid ζ_{31} and ζ_{13} are aggregated together and applied to the operator as the cost of deployed technology degradation. The new version of graph **G** is represented in Fig. 2.

The adjacent matrices for this new setup, which is,

$$\overline{\mathbf{M}} = \begin{array}{c} \\ R \\ I \\ S \end{array} \begin{array}{c} \begin{array}{ccc} R & I & S \end{array} \\ \begin{pmatrix} 0 & 0 & \bar{\zeta}_{13} \\ 0 & 0 & \bar{\zeta}_{23} \\ \bar{\zeta}_{31} & \bar{\zeta}_{32} & 0 \end{pmatrix} \end{array}$$

where,
$\bar{\zeta}_{13} = \{Cost^{-}_{deg}\}$,
$\bar{\zeta}_{23} = \{Bill\ reduction^{+}\}$,
$\bar{\zeta}_{31} = \{\pi^{+}_{DR}, \pi^{-}_{Purchasing}, \pi^{+}_{Selling}\}$,
$\bar{\zeta}_{32} = \{Peak\ shaving^{+}\}$.

According to the relationship between the business model elements, it is possible to calculate the profit of each action. On the other hand, since the energy transaction is the main action of micro-grid elements, we can reformulate the profit function accordingly.

$$Profit(.|\bar{\zeta}_{i,j})(t) = \sum_{i=1}^{N}\sum_{j=1}^{N} u^{i}_{a}\ E^{i}_{a}\ (Cost(.|\bar{\zeta}_{i,j}) - Revenue(.|\bar{\zeta}_{i,j})), \quad (3)$$

where,
$u_a \in \{0, 1\}$,
$E_a \in \{E_G, E_C\}$,
u_a is the binary that shows each element's activating status, E_G is the amount of each element's energy generation, and E_C is the amount of each element's energy consumption. Energy generation is always associated with the cost of fuel and degradation of generators, revealed as $\bar{\zeta}_{13}$ in the business model. On the other hand, the reduction in energy demand will result in bill reduction and benefits for the utility grid, such as peak reduction known in the business model as $\bar{\zeta}_{23}$ and $\bar{\zeta}_{32}$, respectively. $\bar{\zeta}_{31}$ determines the policies and prices that govern all energy transactions between NZR-MG and the utility grid. Hence, we calculate the profit as follows.

$$Profit(.|\bar{\zeta}) = E_G\ Revenue(.|\bar{\zeta}_{23}, \bar{\zeta}_{32}) - E_G\ Cost(.|\bar{\zeta}_{13}) - E_C\ Cost(.|\bar{\zeta}_{31}). \quad (4)$$

The next section describes the revenue and cost associated with each energy transaction.

4. NZR-MG elements and constraints

We assume in our NZR-MG, 130 households are equipped with 400 kW PVs, 600 kWh BESS, and 300 kW fuel cells. This micro-grid is maintained and developed by the third party as the system operator. This NZR-MG trades energy with the Korean electric power company (KEPCO) as a transmission system operator (TSO)/distribution system operator (DSO) at a medium voltage level. In Korea, apartment complexes electrifying with medium-voltage (MV) levels can make single or general power purchase agreements with KEPCO. A single contract determines charges for the apartment complexes based on the total electricity consumption of residents and common areas; however, in a general one, the electricity consumption of each household and shared facilities shall be separately billed. In our study, we considered that there is a single contract for the apartment complex. Therefore, KEPCO installed bi-directional metering in the point of common coupling (PCC). The MV level will supply the public areas, while a private 22.9 kV/0.4 kV transformer to provide a low voltage (LV) line for residential units. Using the sum of energy usage divided

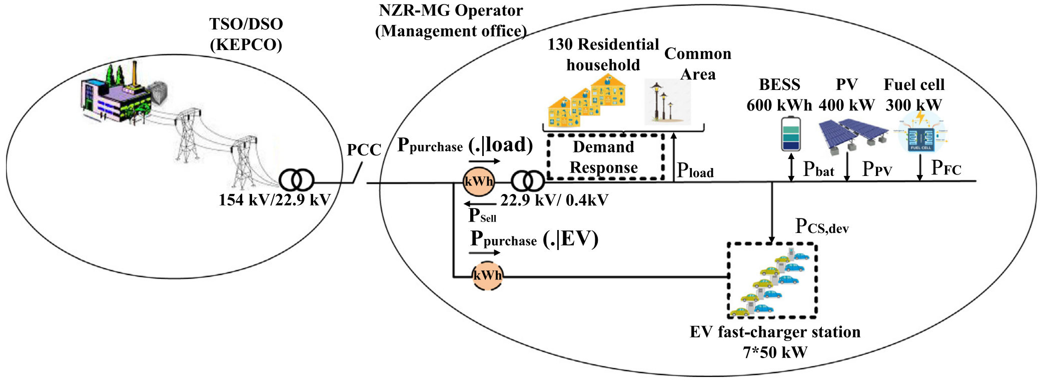


Fig. 3. Existing and proposed elements of understudy NZR-MG.

by the number of households, the management office that has the role of micro-grid operator calculates the electricity bills. To provide a real plan for maximizing all stakeholders profits, the system operator will develop NZR-MG by installing seven fast-charging stations and implementing a demand response schedule based on Korean electricity consumers' behavior. It is noted that the EV charging stations will have their electricity contract from the building and is under the control of the micro-grid operator and try to make maximum profit from selling energy to EVs. Fig. 3 represents the elements of the NZR-MG project in which dotted lines represent the elements of the developed micro-grid. The following sections describe the models used for each NZR-MG component.

4.1. RESs characteristics and constraints

PV is the nature-based generation unit deployed in the understudy NZR-MG. Since PV power output is a random variable, the PV historical data reveal its stochastic behavior. To satisfy the emission-free objective of micro-grid stakeholders and to motivate the minimum dependency of NZR-MG on the utility grid, we ignore the cost of power generation of PV in this study.

$$Cost(P_{PV}|\overline{\zeta}_{13})(t) = 0. \tag{5}$$

Since we consider the PV generation revenue in the energy transaction to the utility grid, the only limitation of PV performance is its minimum and maximum power generation.

$$P^{i}_{PV,min} \le P^{i}_{PV}(t) \le P^{i}_{PV,max}, \qquad 1 \le i \le M, \tag{6}$$

where M represents the number of PVs in the NZR-MG and P_{PV} is PVs output power, respectively.

4.2. Fuel cell characteristics and constraints

The fuel cell is one of the power generation sources of our NZR-MG that serves demand in the PV absence. The current study ignores the cost of fuel cell degradation and applies costs for fuel cell energy production based on fuel price.

$$Cost(FC|\overline{\zeta}_{13})(t) = (P_{FC}(t)\ Cost_{Fuel}(t))/\eta_{FC} \qquad , \tag{7}$$

where P_{FC} is the output power of the fuel cell and η_{FC} is its efficiency. With the same approach as PV, the revenue from fuel cell performance is considered in the power transaction with the utility grid. Additionally, the fuel cell state can change in a certain time step concerning its minimum up and down time characteristics as follows.

$$(T^{on}_{FC}(t-1) - T^{up}_{FC})(u_{FC}(t-1) - u_{FC}(t)) \ge 0 \tag{8}$$

$$(T^{off}_{FC}(t-1) - T^{down}_{FC})(u_{FC}(t) - u_{FC}(t-1)) \ge 0 \tag{9}$$

where T^{up}_{FC} and T^{down}_{FC} determine the fuel cell's minimum up and down time and T^{on}_{FC} and T^{off}_{FC} are fuel cell's duration of being on or off, respectively. u_{FC} is a binary value that shows whether the fuel cell is on or off.

4.3. BESS modeling and constraints

BESS controls the stochastic characteristic of RESs in our NZR-MG. We indicate the SoC of BESS each time by:

$$SoC(t) = SoC(t-1) + \frac{P_{bat}\ \Delta t}{E_b\ \eta_{bat}}, \tag{10}$$

where P_{bat} has a positive value in the charging state and negative amount in the discharging mode, E_b is the reference capacity of the battery, Δt is the time slot of charging/discharging of the battery, and η_{bat} is the battery charging and discharging efficiency. However, to protect the battery capacity from deterioration we limit the SoC variation and battery power according to (11), (12).

$$SoC_{min} \le SoC(t) \le SoC_{max}, \tag{11}$$

$$P_{bat,min} \le P_{bat}(t) \le P_{bat,max}. \tag{12}$$

In this study, we estimate the cost of battery degradation according to (13) [21] and consider the revenue from battery power provision in energy transactions with the utility grid.

$$Cost(BESS|\overline{\zeta}_{13})(t) = \alpha\ (SoC(t) - SoC(t-1))^2. \tag{13}$$

From [21], we assigned 0.9 to the battery degradation coefficient α proportional to our BESS capacity and other NZR-MG elements specifications to control the BESS number of charging and discharging cycles.

4.4. Demand response modeling

4.4.1. Residential load modeling and demand response

The under-study NZR-MG consists of three residential building blocks located in Seoul. We considered the residential load profile of the NZR-MG according to Korean residential power consumers' behavior. The details of the monthly load profile estimation and hired databases are represented in Appendix C.

We did not apply demand response in the base case model. To grant the accuracy of our business model considering profit for each stakeholder, we added demand response to the developed NZR-MG model. In Korea, the electricity market operator, Korean power exchange company (KPX), determines the system marginal price (SMP) for trading power in the electricity market. In addition, KPX estimates power shortage and issue biding commands through aggregators for demand response resources.

The period of applying demand response according to the KPX policy should not exceed 60 h [22]. We categorize loads within demand response into shiftable, reducible, and non-interruptible. In this paper, we applied demand response to heating and cooling system as reducible loads. The reasons for this arrangement are twofold. According to the understudy NZR-MG electricity contract with KEPCO, the energy price is a three-stage progressive rate ($Price^{t,d}_{3sp}$) determined in Section 4.5. Since the sum of the load consumption during the day determines the cost of power, shifting the load is not an incentive. In addition, most of the electricity consumed by residential units goes to heating and cooling systems. The energy reduction is implemented by controlling the indoor temperature (θ_{in}) in a range of residents desirable (θ_{min}, θ_{max}). Participation probability for each demand response unit in every event day ($\mathcal{P}_{DR}$) is calculated with a Gaussian mixture model based on Korean electricity market records [23]. Hence, we deployed the following equation to apply demand response to our model and calculate our NZR-MG revenue from this action.

$$Revenue(DR|\overline{\zeta}_{31},\overline{\zeta}_{23})(t) = P_{DR,dev}(t)\mathcal{P}(P_{DR,dev}|t)Price^{t,d}_{3sp} - u_{ax}\Delta\theta^{t}, \tag{14}$$

where,

$$\Delta\theta^{t} = \begin{cases} (\theta_{min} - \theta^{t}_{in}), & if\ \theta^{t}_{in} < \theta_{min}, \\ (\theta^{t}_{in} - \theta_{max}), & if\ \theta^{t}_{in} > \theta_{max}, \\ 0, & otherwise, \end{cases} \tag{15}$$

$$\mathcal{P}(P_{DR,dev}|t) = 0.35exp(1.69(t-10)^2)+0.25exp(25(t-1)^2), \tag{16}$$

u_{ax} is the anxiety coefficient of consumers from undesired indoor temperature, and t is the time of demand response event. The outdoor temperature historical data and heating and cooling system power usage ($P^{H\&C}(t)$) facilitate the indoor temperature calculation according to (17) [24].

$$\theta^{t}_{in} = \theta^{t-1}_{in} + \mathcal{K}_1(\theta^{t-1}_{out} - \theta^{t-1}_{in}) + \mathcal{K}_2 P^{H\&C}(t), \tag{17}$$

where $\mathcal{K}_1$ and $\mathcal{K}_2$ are coefficients to determine indoor temperature. The power consumption of the heating and cooling system and indoor temperature amount is limited as follows.

$$\theta_{min} \le \theta^{t}_{in} \le \theta_{max}, \tag{18}$$

$$P^{H\&C}_{min} \le P^{H\&C}(t) \le P^{H\&C}_{max}. \tag{19}$$

The amount of reduction in power consumption of the heating and cooling system will determine the amount of demand response power according to (20), (21).

$$\Delta P^{H\&C} = P^{H\&C}(t) - P^{H\&C}(t-1), \tag{20}$$

$$P_{DR,dev} = \Delta P^{H\&C}. \tag{21}$$

4.4.2. EV modeling

Determination of EV power demand depends on the owner's driving habit and the charging characteristics of the EV. In this paper, concerning EV battery lifetime, we consider SoC between 0.2 and 0.8. Since the charging station is a fast charger, the power demand during this range of SoC stays almost constant [25]. The arrival time of EV in this paper assumes that have Gaussian distribution according to (22).

$$\mathcal{P}(t,EV) = \frac{1}{\delta_{tarr}\sqrt{(2\pi)}}exp(-\frac{(t-\mu_{tarr})^2}{2\delta_{tarr}{}^2}), \tag{22}$$

where $\mathcal{P}(t,EV)$ is the possibility of EV arriving home at time t, δ_{tarr} is the standard deviation, and μ_{tarr} is the average value and equal to 38.8 km and 21.9 km for personal purpose travels, respectively, and according to the Korean private vehicles travel pattern [26]. However, we consider each charging pile is available to serve EVs the whole time. Hence, we model the demand of the charging station in each time based on the home arrival time of EVs as follows [27].

$$P_{CS,dev}(t) = \sum_{n=1}^{k} P_{n,rated}\mathcal{P}(t,EV), \tag{23}$$

where k is the number of charging piles at the charging station and $P_{n,rated}$ is the rated power of charging piles. Since the charging stations in our model are fast DC chargers, we did not apply the vehicle-to-grid (V2G) in our model [28]. The price of power purchasing for EV is according to Table B.1 of Appendix B [29].

Since the NZR-MG operator objective is the profit maximization from selling power to the EVs, demand response is not applied to EV load. However, to respect the interests of the utility company, the NZR-MG operator firstly supplies EVs by surplus power of RESs if the residential loads are already fulfilled. The operator sells the power to EV owners slightly higher than the prices in Table B.1 of Appendix B when the utility grid supplies EVs. Therefore, the revenue from the EV charging station is calculated as follows.

$$Revenue(EV|\overline{\zeta}_{31})(t) = P_{CS,dev}(t)\ Price^{t,d}_{EV}, \tag{24}$$

where,

$$Price^{t,d}_{EV} = \rho Price^{KEPCO}_{EV}. \tag{25}$$

Table B.1 of Appendix B determines the $Price^{KEPCO}_{EV}$, and we assume ρ is equal to 1 if the NZR-MG supplies $P_{CS,dev}(t)$ and 1.1 if the $P_{CS,dev}(t)$ is purchased from the utility grid and determined by $P_{purchase}(.|EV)(t)$.

4.5. The utility grid modeling

Our NZR-MG in the PCC purchases power from the grid to compensate for power shortages in the absence of RESs and when the cost of power provision from other resources is higher than purchasing power from the utility grid. Table B.2 of Appendix B shows the price of power purchasing from the grid for residential consumption.

In addition, the NZR-MG can participate in the electricity market to sell its excess power based on SMP determined by KPX [30]. The NZR-MG elements should coordinate to respect the balance between generation and consumption. We consider RESs as the high prior sources of power generation and EV as the load with the profile according to Section 4.4.2 that should be served continuously. The constraints for trading power with the utility grid and power balance in the NZR-MG are as follows.

$$P_{net}(t) = P_{PV}(t) - (P_{load}(t) - P_{DR,dev}(t)), \tag{26}$$

$$\begin{aligned} &P_{net}(t) + P_{FC}(t) + P_{bat}(t) + P_{purchase}(.|load)(t) + P_{purchase}(.|EV)(t) \\ &\quad -P_{sell}(t) - P_{CS,dev}(t) = \lambda, \quad \forall t \quad \lambda = 0, \end{aligned} \tag{27}$$

$$P_{purchase}(.|load)(t)P_{sell}(t) = 0, \tag{28}$$

where $P_{purchase}(.|load)(t)$ is the amount of power that is purchased from the utility grid to supply load, $P_{purchase}(.|EV)(t)$ is the amount of power that is purchased from the utility grid to supply EVs, and $P_{sell}(t)$ is the amount of sold power to the grid at each time t. The NZR-MG is not allowed to sell and purchase power simultaneously, as shown in (28). The revenue and cost of power trading with the utility grid are as follows.

$$Revenue(P_{sell}|\overline{\zeta}_{31})(t) = P_{sell}(t)SMP^{t,d}, \tag{29}$$

$$Cost(P_{purchase}|\overline{\zeta}_{31})(t) = P_{Purchase}(.|load,t)Price^{t,d}_{3sp}. \tag{30}$$

4.6. NZR-MG profit modeling

According to the proposed profit function in (4) and the cost function of the NZR-MG elements represented in the previous sections, we define the objective function for our NZR-MG as follows.

$$
\begin{aligned}
&Max\ Profit(t) = \\
&Max\left[\ Revenue\left(P_{sell}(t) + P_{DR,dev}(t) + P_{CS,dev}(t)\right) - \right. \\
&\left. Cost\left(P_{Purchase}(.|load,t) + P_{bat}(t) + P_{FC}(t)\right)\ \right],
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
&Max\ Profit(t) = \\
&Max\left[\left(\ P_{sell}(t)SMP^{t,d} + (P_{DR,dev}(t)\mathcal{P}(P_{DR,dev}|t)Price^{t,d}_{3sp} - u_{ax}\Delta\theta^{t}) + \right.\right. \\
&\left. P_{Purchase}(.|EV,t)Price^{t,d}_{EV}\ \right) - \left(\ P_{Purchase}(.|load,t)Price^{t,d}_{3sp} + \right. \\
&\left.\left. \alpha(SoC(t) - SoC(t-1))^2 + (P_{FC}(t)Cost_{Fuel})/\eta_{FC}\ \right)\right]
\end{aligned}
\tag{32}
$$

subject to
(6), (8), (9), (11), (12), (15), (18), (19), (27), (28).

5. DRL solution algorithms

The complexity and the high-dimensional problem of NZR-MG encourage deploying model-free DRL to solve micro-grid EMS problems. Our baseline solutions encompass three DRL algorithms, DQN, DDQN, and DDPG. To provide an accurate and fast converged solution, DDQN combined with MINLP and TL. The first step in solving RL is the Markov decision process (MDP) arrangement. Therefore, this section discusses the MDP arrangement and the solution algorithms for our understudy NZR-MG.

5.1. NZR-MG MDP arrangement

MDP is a 4-tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}\}$, where $\mathcal{S}$ is the set of the environment states, $\mathcal{A}$ is the agent's actions set, $\mathcal{T}$ is the transition function that shows the probability of transferring to the next state s_{t+1} while agent takes action a_t in time t, and $\mathcal{R}$ is the set of rewards that agent assigns to each action.

5.1.1. States

We arranged our understudy NZR-MG elements according to Fig. 3 with limitations and constraints specified in Section 4. Since the objective of NZR-MG advent is spreading RESs deployment, their sufficiency in supplying loads, P_{net}, calculated by (26), is one of the most critical states of the environment. The other states are other generators' status, including fuel cell output power (P_{FC}) and available power of BESS in discharging mode (P_{bat}) and its SoC. Because NZR-MG can trade energy with the utility grid, selling surplus power to the utility grid price ($SMP^{t,d}$) and purchasing energy from the grid rate, $Price^{t,d}_{3sp}$ based on Table B.2 of Appendix B, are the other states. To distinguish the $Price^{t,d}_{3sp}$, we need to calculate the sum of electrical power NZR-MG purchases from the grid to supply residential loads each day, as shown in (33).

$$
SUM_{Pur,load} = \sum_{t=1}^{T} P_{Purchase}(.|load, T),
\tag{33}
$$

where T is the number of the time period that the meter records the electrical power usage in each day.

For the developed case, the EV charging station power consumption ($P_{CS,dev}$) and charging price, $Price^{t,d}_{EV}$ based on Table B.1 of Appendix B, will be added to the state space. In addition to that, in the developed case to contribute NZR-MG in the demand response program, according to Section 4.4.1, the inside and outside building temperature ($\theta_{in}(t)$, $\theta_{out}(t)$) are the other states of the environment.

Therefore, we specify the state space by

$$
\begin{aligned}
\mathcal{S} = \{&P_{net}(t), P_{CS,dev}(t), SoC, SUM_{Pur,load}, \\
&SMP^{t,d}, Price^{t,d}_{3sp}, Price^{t,d}_{EV}, T^{on}_{FC}, T^{off}_{FC}, \theta_{in}(t), \theta_{out}(t), time\}.
\end{aligned}
\tag{34}
$$

5.1.2. Actions

Our NZR-MG includes controllable and stochastic resources of energy. We deployed historical data to predict stochastic characteristics of RESs and loads. The agent can control the other elements, including BESS, fuel cell, and the amount of trading power from the utility grid. Hence, the action space $\mathcal{A}$ is defined by:

$$
\mathcal{A} = \{\mathcal{A}_{BESS}, \mathcal{A}_{FC}, \mathcal{A}_{sell}, \mathcal{A}_{purchase|load}, \mathcal{A}_{purchase,dev|EV}, \mathcal{A}_{DR,dev}\}.
\tag{35}
$$

Each action can follow a discrete or continuous space. We present in this paper several DQN-based algorithms and DDPG procedures for managing energy transactions in the micro-grid, which have different approaches related to the action space. The DQN-based agent selects actions from a discrete action space, while the DDPG agent takes continuous actions. To determine the discrete action, we discrete battery and fuel cell power according to (36) and (37), respectively, and arrange discrete action space with a mixture of these actions and other elements of action space as follows.

$$
P_{bat,disc} = \{-200, -150, -100, -50, 0, 50, 100, 150, 200\},
\tag{36}
$$

$$
P_{FC,disc} = \{0, 100, 200, 300\},
\tag{37}
$$

$$
\begin{aligned}
\mathcal{A}_{disc} = \{&P_{bat,disc}, U_{FC}, P_{FC,disc}, P_{sell}, \\
&P_{purchase}(.|load), P_{purchase,dev}(.|EV), \\
&\mathcal{A}_{purchase,dev|EV}, \mathcal{A}_{DR,dev}\},
\end{aligned}
\tag{38}
$$

where,
$\mathcal{A}_{purchase,dev|EV} = \{0, 1\}$.
$\mathcal{A}_{DR,dev} = \{0, 1\}$.
$P_{bat,disc}$: The BESS amount of charging or discharging per kW, according to the predetermined range.
U_{FC}: The fuel cell state of being on or off.
$P_{FC,disc}$: The fuel cell amount of charging or discharging per kW, according to the predetermined range.
P_{sell}: The amount of selling power to the utility grid in the PCC per kW.
$P_{purchase}(.|load)$: The amount of buying power from the utility grid in the PCC per kW.
$P_{purchase,dev}(.|EV)$: The amount of power purchased to supply EV from the utility grid or NZR-MG per kW in the developed case.
$\mathcal{A}_{purchase,dev|EV}$: The state of supplying EV from the utility grid or from NZR-MG resources in the developed NZR-MG.
$\mathcal{A}_{DR,dev}$: The state of reducible load participation in the demand response plan in the developed NZR-MG.

The continuous action space is defined as follows.

$$
\begin{aligned}
\mathcal{A}_{cont} = \{&P_{bat}(t), P_{FC}(t), P_{sell}(t), P_{purchase}(.|load)(t), \\
&P_{purchase,dev}(.|EV)(t), P_{DR,dev}(t)\},
\end{aligned}
\tag{39}
$$

where,
P_{bat}: The amount of BESS charging or discharging per kW.
P_{FC}: The amount of power generated by the fuel cell per kW.

P_{sell}: The amount of selling power to the utility grid in the PCC per kW.
$P_{purchase}(.|load)$: The amount of buying power from the utility grid in the PCC per kW.
$P_{purchase,dev}(.|EV)$: The amount of power purchased to supply EV from the utility grid or NZR-MG in the developed case.
$P_{DR,dev}$: The amount of load reduction by participation in demand response per kW.

5.1.3. Transition function

In our understudy NZR-MG, BESS is the only element we can estimate its next state with (10). The other elements' status, such as RESs, Loads, EVs, and SMP, is unknown. Due to this, we utilize DRL, which is model free technique and can support the uncertain environment of NZR-MGs. To provide an accurate next state estimation, we start from a stochastically choosing SoC at the starting point of each episode and determine RESs, Loads, EVs, and SMP present states from the historical data. While the next step of P_{bat} is determined according to (10), the agent initially chooses the other elements in a stochastic manner. This random selection will be conducted to the best action selection during the agent's learning process by assigning the reward to each action, which is the nature of the DRL approach to solving problems.

5.1.4. Reward

Given that we selected two different actions for the DRL, it is appropriate for the reward function to be set based on those actions. However, since we deployed two different approaches in action selection for the DRL agent, we need to modify this reward function accordingly. Therefore, we define the reward function as follows.

$$\mathcal{R} = \left(\sum_{i=1}^{N} \alpha^i \mathcal{A}^i_{dis/cont}(Revenue^i_A - Cost^i_A)\right) + \mathcal{R}_{SoC} + \mathcal{R}_{balance}, \quad (40)$$

where i is the number of each NZR-MG element, $\mathcal{A}_{dis/cont}$ are continuous and discrete actions defined in (38) and (39), and the coefficient α is defined according to the priority of resources in energy provision and determined as follows.

$$\alpha_{PV} \gg \alpha_{bat} \gg \alpha_{FC} > \alpha_{DR,dev} > \alpha_{Trading}. \quad (41)$$

$\mathcal{R}_{SoC}$ is a punishment to keep the SoC in a range defined in (11), and $\mathcal{R}_{balance}$ preserves NZR-MG power balance evaluated by (26) as follows.

$$\mathcal{R}_{SoC} = -2 \times 10^7, \quad if \quad (SoC < SoC_{min}) \quad \| \quad (SoC > SoC_{max}), \quad (42)$$

$$\mathcal{R}_{balance} = -2 \times 10^7, \quad if \quad \lambda \neq 0. \quad (43)$$

5.2. DQN

DQN is an efficient RL algorithm to overcome the curse of dimensionality weakness of Q-learning when solving problems involving large numbers of states, such as EMS in micro-grids [31]. A DQN agent provides environment states as inputs to the DNN, also known as the Q network, in each time step and receives the Q-value of each action from the DNN. The agent selects an action with a higher value based on the ϵ-greedy policy. This process will provide a transition tuple for each time step (s_t, a_t, r_t, s_{t+1}) saved in an experienced buffer. The batch of experience from the replay buffer after each episode termination applies to the Q-network to prevent correlation between inputs for the DNN. To update the weight of the network, the agent deploys a gradient descent approach by estimating the loss function according to (44), which is the mean square error of the difference between the target value y_t, derived from the Bellman equation represented in (45), and the Q-network predicted value as follows.

$$Loss(\theta_t) = \mathbb{E}[(y_t - Q(s_t, a_t|\theta_t))^2], \quad (44)$$

where,

$$y_t = r_t + \gamma maxQ(s_{t+1}, a_t|\theta_t), \quad (45)$$

γ is discounting factor and $maxQ(s_{t+1}, a_t)$ is the maximum future Q-value of the next state.

Algorithm 1 represents the pseudo-code of DQN to solve the EMS problem of our NZR-MG environment.

Algorithm 1: DQN algorithm pseudo-code

```
Initialize $Q_\theta$ and empty replay buffer $\mathcal{D}$ ;
for episode = 1 to E do
    Generate $s_t$ based on (34) ;
    for $t = 1$ to $T$ do
        Generate $a_t$ based on (38) using $\epsilon$-greedy;
        Calculate reward $r_t$ according to (40);
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in $\mathcal{D}$;
        Sample minibatch of transitions;
        if episode terminates at step $t + 1$ then
            $y_t \leftarrow r_t$;
        else
            $y_t \leftarrow (r_t + \gamma maxQ(s_{t+1}, a_t|\theta_t))$;
        Calculate loss function according to (44);
```

5.3. DDQN

The DQN agent's main drawback is that it overestimates the Q-value since it uses the maximum Q-value of all possible actions to update the Q-value. Attaching target network with network parameter of θ' in DDQN updates (44) with the estimation of target value from this network as follows [32].

$$Loss(\theta_t) = \mathbb{E}[(r_t + \gamma maxQ(s_{t+1}, a_t|\theta'_t) - Q(s_t, a_t|\theta_t))^2]. \quad (46)$$

The target network is a copy of the main Q-network frozen for a while to prevent oscillation of the Q-network by estimating its weight from estimation. Algorithm 2 is the DDQN pseudo-code for scheduling EMS of our understudy NZR-MG.

Algorithm 2: DDQN algorithm pseudo-code

```
Initialize $Q^\theta$, $Q^{\theta'} \leftarrow Q^\theta$, and empty replay buffer $\mathcal{D}$ ;
for episode = 1 to E do
    Generate $s_t$ based on (34) ;
    for $t = 1$ to $T$ do
        Generate $a_t$ based on (38) using $\epsilon$-greedy ;
        Calculate reward $r_t$ according to (40);
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in $\mathcal{D}$;
        Sample minibatch of transitions;
        if episode terminates at step $t + 1$ then
            $y_t \leftarrow r_t$;
        else
            $y_t \leftarrow (r_t + \gamma maxQ(s_{t+1}, a_t|\theta'_t))$;
        Calculate loss function according to (46);
        Update $Q_\theta$ with SGD by minimizing loss function;
        Set Target network weights after several steps by
          $Q^{\theta'} \leftarrow Q^\theta$;
```

5.4. DDPG

As with DQN, DNNs estimate the DDPG value function. However, DDPG is an actor–critic method and requires another DNN to approximate the strategy function. DDPG also benefits from replay buffer and fixed Q target network tricks of DQN. Therefore, there are four DNNs in DDPG, including θ^{Q} and $\theta^{Q'}$ for online and target value function estimation and θ^{μ} and $\theta^{Q^{\mu'}}$ for online and target strategy function evaluation, respectively [33]. The online value network is updated similarly to the DQN, whereas the online policy network is updated based on (47).

$$\theta^{\mu} \leftarrow \theta^{\mu} - \alpha_c \nabla_{\theta^{\mu}} J, \tag{47}$$

where α_c is the policy network learning rate, and $\nabla_{\theta^{\mu}} J$ is the gradient of the objective function to estimate the maximum of the objective function. $\nabla_{\theta^{\mu}} J$ is equivalent to the action-value function expected gradient and calculated as follows.

$$\begin{aligned} &\nabla_{\theta^{\mu}} \mathcal{J}(\theta) \\ &\approx \frac{1}{N} \sum_{i} [\nabla_a Q(s, a|\theta^{Q}, s = s_i, a = \mu(\mathcal{S}_i)) \, \nabla_{\theta^{\mu}} \, \mu(\mathcal{S}|\theta^{\mu}, s = s_i)]. \end{aligned} \tag{48}$$

DDPG updates actor and critic target networks with a soft update mechanism to provide stability in learning process, according to (49), (50).

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}, \quad \tau \ll 1, \tag{49}$$

$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}, \quad \tau \ll 1. \tag{50}$$

where τ is the target smoothing factor set to 0.005 in this study.

In the utilized DDPG algorithm in this paper, we also used mean-zero Gaussian noise to improve exploration in action selection.

Algorithm 3: DDPG algorithm pseudo-code

```
Initialize $Q^{\theta'} \leftarrow Q^{\theta}$ , $Q^{\mu'} \leftarrow Q^{\mu}$, and empty replay buffer $\mathcal{D}$
 ;
for episode = 1 to E do
    Initialize random process $\mathcal{N}$ ;
    Generate $s_t$ based on (34) ;
    for $t = 1$ to T do
        Generate $a_t$ based on (39) using current policy and
         exploration noise ($a_t = \mu(S|\theta^{\mu}) + \mathcal{N}_t$);
        Calculate reward $r_t$ according to (40);
        Store transition ($s_t, a_t, r_t, s_{t+1}$) in $\mathcal{D}$;
        Sample minibatch of transitions;
        if episode terminates at step $t + 1$ then
            $y_t \leftarrow r_t$;
        else
            $y_t \leftarrow (r_t + \gamma max Q(s_{t+1}, a_t|\theta'_t))$;
        Update critic network by minimizing loss function
         (46);
        Update policy network using sampled policy
         gradient according to (47);
        Update critic and policy target networks by Setting
         $\theta'$ and $\mu'$ according to (49), (50);
```

5.5. Proposed DDQN algorithm with MINLP and TL contribution

The objective of the RL method is to maximize the cumulative reward over time. By selecting actions with the highest value function, DQN, DDQN, and DDPG accomplish this objective. Accordingly, the quality of the Q-value function is a determinant of learning speed and accuracy. Deploying the maximum value to estimate future rewards makes all Q-learning methods subjected to over-estimation. We combine DDQN and MINLP to prevent overestimation and accelerate the optimization process to converge to the accurate results in this paper . In the deployed DQN, DDQN, and DDPG, which is a pure DRL method, the definition of the reward function according to the exact profit represented in (32) is not viable and defined according to (40). As with all RL methods, the reward function will imitate the original cost function. On the other hand, DDPG has a continuous action space, and its combination with the MINLP solver encounters a large amount of action resulting in failover in action evaluation. As a result, DDQN is combined with MINLP solvers to provide agents with more effective rewards. Instead of discrete action space (38), MINLP+DDQN determines several strategies with upper and lower bounds of P_{bat} as follows.

$$\begin{aligned} Strategy_{MINLP+DDQN} = \{&[-200, -150], [-150, -100], \\ &[-100, -50], \\ &[-50, 0], [0, 50], [50, 100], [100, 150], [150, 200]\}. \end{aligned} \tag{51}$$

These strategies with the other elements constraints provide upper and lower bounds of MINLP solver according to (52), (53).

$$\begin{aligned} Ub = [&P_{bat,min}, P_{FC,min}, P_{sell,min}, P_{purchase|load,min}, \\ &P_{purchase,dev|EV,min}, P_{DR,dev,min}], \end{aligned} \tag{52}$$

$$\begin{aligned} Lb = [&P_{bat,max}, P_{FC,max}, P_{sell,max}, P_{purchase|load,max}, \\ &P_{purchase,dev|EV,max}, P_{DR,dev,max}], \end{aligned} \tag{53}$$

where P_{bat} is a member of (51).

Algorithm 4: TL+DDQN+MINLP approach

```
For subtask 1 initialize $Q^{\theta}$, $Q^{\theta'} \leftarrow Q^{\theta}$ randomly;
For subtask 2 initialize $Q^{\theta}$, $Q^{\theta'} \leftarrow Q^{\theta}$ from subtask 1
 learned parameters;
For subtask 3 initialize $Q^{\theta}$, $Q^{\theta'} \leftarrow Q^{\theta}$ from subtask 2
 learned parameters;
Initialize an empty replay buffer $\mathcal{D}$;
for episode = 1 to E do
    Generate $s_t$ based on (34) ;
    for $t = 1$ to T do
        For subtask 1 and subtask 2 generate $a_t$ based on
         (38) using $\epsilon$-greedy ;
        For subtask 3 Generate $a_t$ based on (51) using
         $\epsilon$-greedy ;
              run MINLP solver for (32) with applying
         related constraints;
        Deliver optimal actions to DDQN agent;
        For subtask 1 calculate reward according to (54);
        For subtask 2 calculate reward (55);
        For subtask 3 calculate reward $r_t$ according to (40);
        Store transition ($s_t, a_t, r_t, s_{t+1}$) in $\mathcal{D}$;
        Sample minibatch of transitions;
        if episode terminates at step $t + 1$ then
            $y_t \leftarrow r_t$;
        else
            $y_t \leftarrow (r_t + \gamma max Q(s_{t+1}, a_t|\theta'_t))$;
        Calculate loss function according to (42);
        Set Target network weights after several steps by
         $Q^{\theta'} \leftarrow Q^{\theta}$;
```

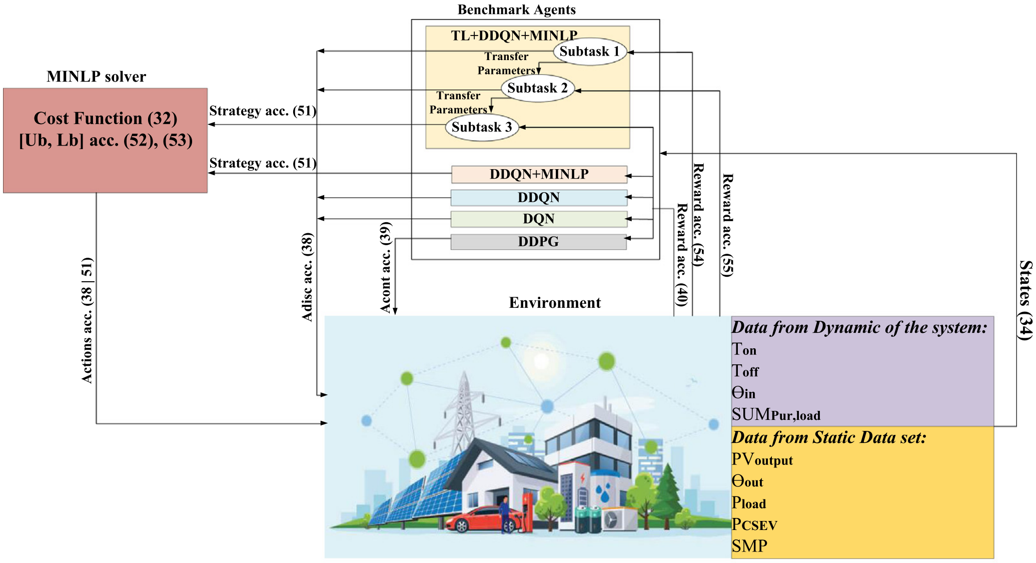
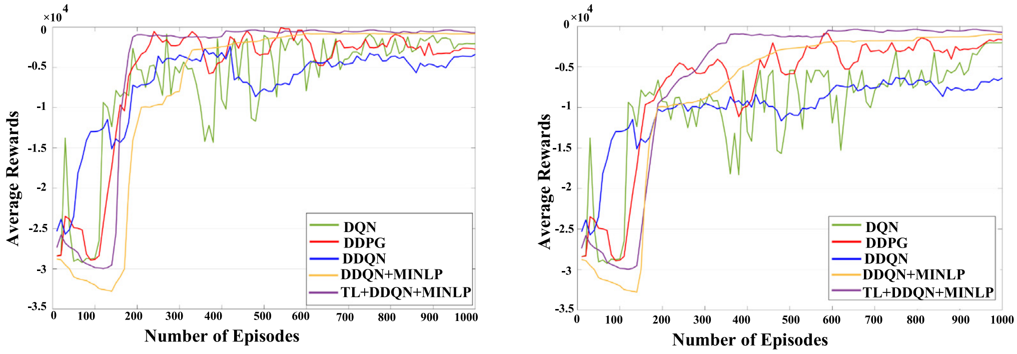


Fig. 4. Our NZR-MG EMS optimization approach flowchart.



(a) DQN, DDQN, DDPG , DDQN+MINLP, and TL+DDQN+MINLP training process comparison for base case

(b) DQN, DDQN, DDPG , DDQN+MINLP, and TL+DDQN+MINLP training process comparison for developed case

Fig. 5. Solution algorithms training process comparison in base case and developed case.

However, MINLP computation costs will subject the learning process to delay when finding optimized actions. To overcome this weak point, we utilized TL techniques. This approach accelerates the learning process of the agent. Transfer learning involves taking what has been learned on one task and applying it to another possibly related challenge. To this end, we divided NZR-MG optimization into three subtasks. In subtask 1, the agent will justify network weights to prevent extreme behavior, related to SoC and Power balance limitations. Subtask 1 reward function is calculated by (54). Since the agent should consider the cumulative power purchase from the utility grid to avoid falling price into the higher stages, subtask 2 is an effort to determine network parameters concerning this limitation. Subtask 2 actions will be evaluated with reward function according to (55). The third subtask is considering the environment with whole objectives and constraints. Algorithm 4 reveals pseudo-code of TL+MINLP+DDQN, and Fig. 4 represents a flowchart of different approaches hired in this paper to schedule EMS for NZR-MG.

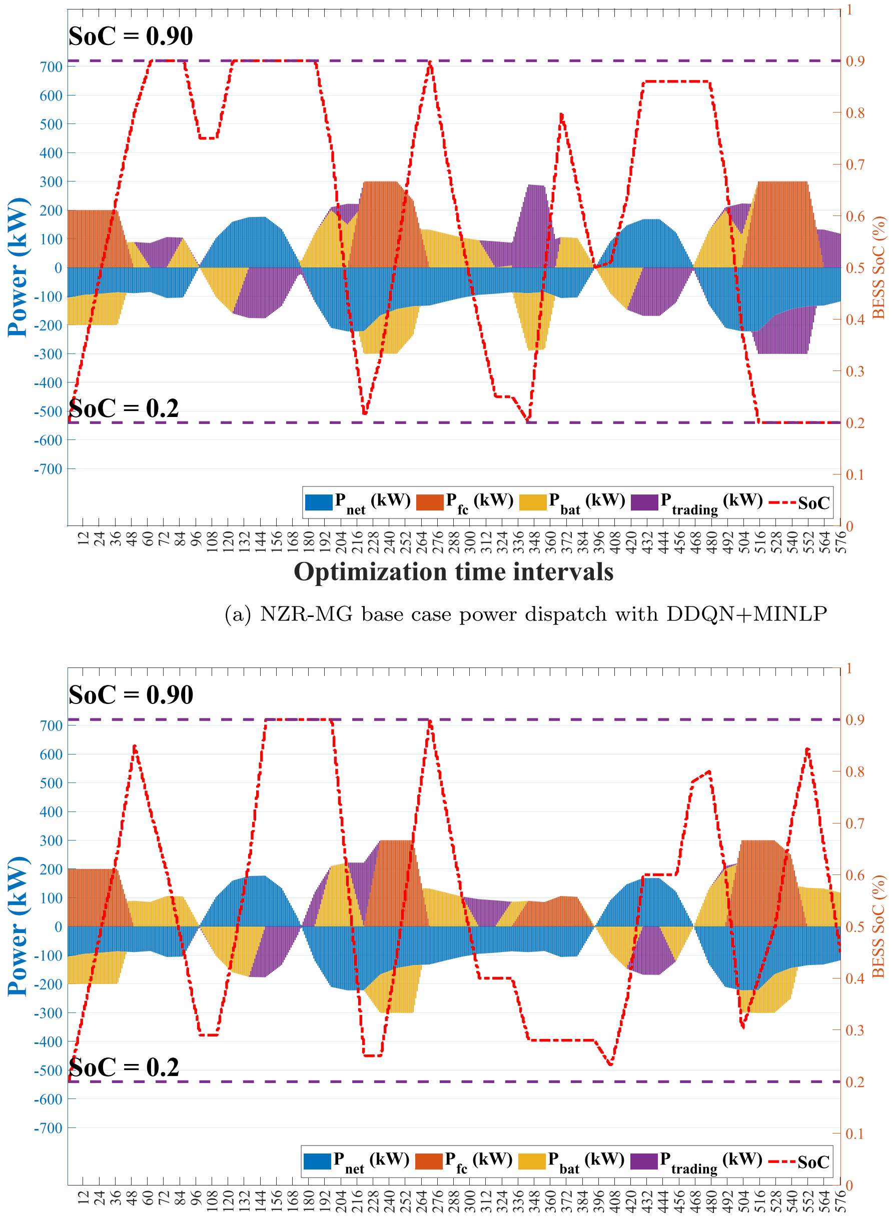
$$r_t = Reward_{SoC} + Reward_{balance}, \tag{54}$$

$$r_t = \alpha_{Trading} \mathcal{A}_{dis}(Revenue_A - Cost_A) \tag{55}$$

6. Results and discussion

6.1. Experimental setup

We evaluate our proposed methodology by planning EMS for the base and developed cases of NZR-MG, introduced in Section 4. Table D.1 shows the NZR-MG elements' specifications and constraints. The annual historical data of PV output power for a one-hour time slot according to the weather condition of Seoul is provided from [34]. The load profile follows the pattern represented in Section 4.4.1. NZR-MG base case consists of PV output power, average load profile, hourly SMP, and $Price^{t,d}_{3sp}$ of each selected month while adding the inside and outside temperature, EV load profile, and $Price_{EV}$ provide the environment for developed NZR-MG specified in Section 4.4.2. As part of setting up our NZR-MG environment, we divided our historical data into training and testing sets. Training data consists of 243 days, including two months from each season, and testing data consists of 122 days remaining in the year. We schedule our NZR-MG for every 5 min. To this end, for both base and developed cases in the training process, the number of sampling for each episode is 69,984, which is the number of training days times the number



(a) NZR-MG base case power dispatch with DDQN+MINLP

(b) NZR-MG base case power dispatch with TL+DDQN+MINLP

Fig. 6. Comparison of NZR-MG basecase hourly power dispatch with DDQN+MINLP and TL+DDQN+MINLP.

of 5-min periods in a day. The test process also samples actions 35,136 times, according to the test data size.

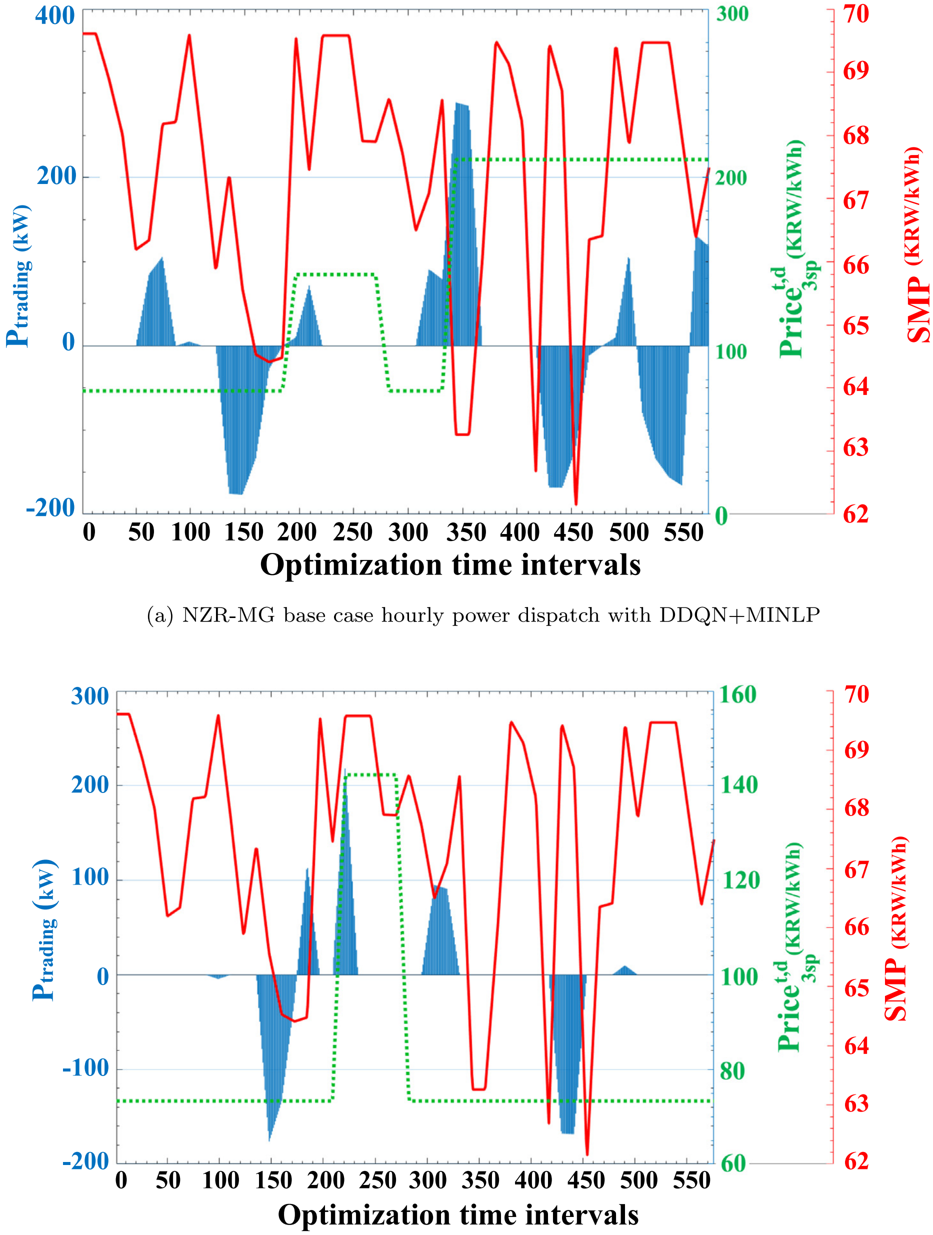
We deployed DQN, DDQN, DDPG, MINLP+DDQN, and TL+MINLP+ DDQN algorithms to schedule energy resources for the base and developed cases of NZR-MG. Table D.1 of Appendix D outlines the hyper-parameters for each algorithm. Q-networks and policy networks have two fully connected hidden layers. Each hidden layer has 200 ReLU neurons. We fix the replay buffer capacity to 5000, and in each gradient descent step, the minibatch size of samples is 256. The algorithms trained over 10,000 episodes.

For DQN, DDQN, and MINLP-DDQN, action selection follows the ϵ-greedy policy. Actions are randomly selected in the first 100 episodes to explore the state–action space as effectively as feasible. The next step is to choose actions using the ϵ-greedy policy according to (56). As can be seen, in (56), ϵ-greedy is implemented in DDPG by adding noise with a zero Gaussian distribution pattern.

$$a_t = \begin{cases} Any \quad a_t, & probability \quad \varepsilon \\ maxQ_t(a) & probability \quad 1-\varepsilon, \quad if \begin{cases} DQN \\ DDQN \\ DDQN + MINLP \\ TL + DDQN + MINLP \end{cases} \\ \mu(\mathcal{S}_t|\theta^{\mu}) + \mathcal{N}_t, & probability \quad 1-\varepsilon, \quad if \quad DDPG \end{cases} \tag{56}$$

6.2. Experimental results

This section represents a comprehensive comparison of each algorithm in planning EMS of NZR-MG. We conduct experiments on MATLAB Simulink (2022a) and test simulations on a machine



(a) NZR-MG base case hourly power dispatch with DDQN+MINLP

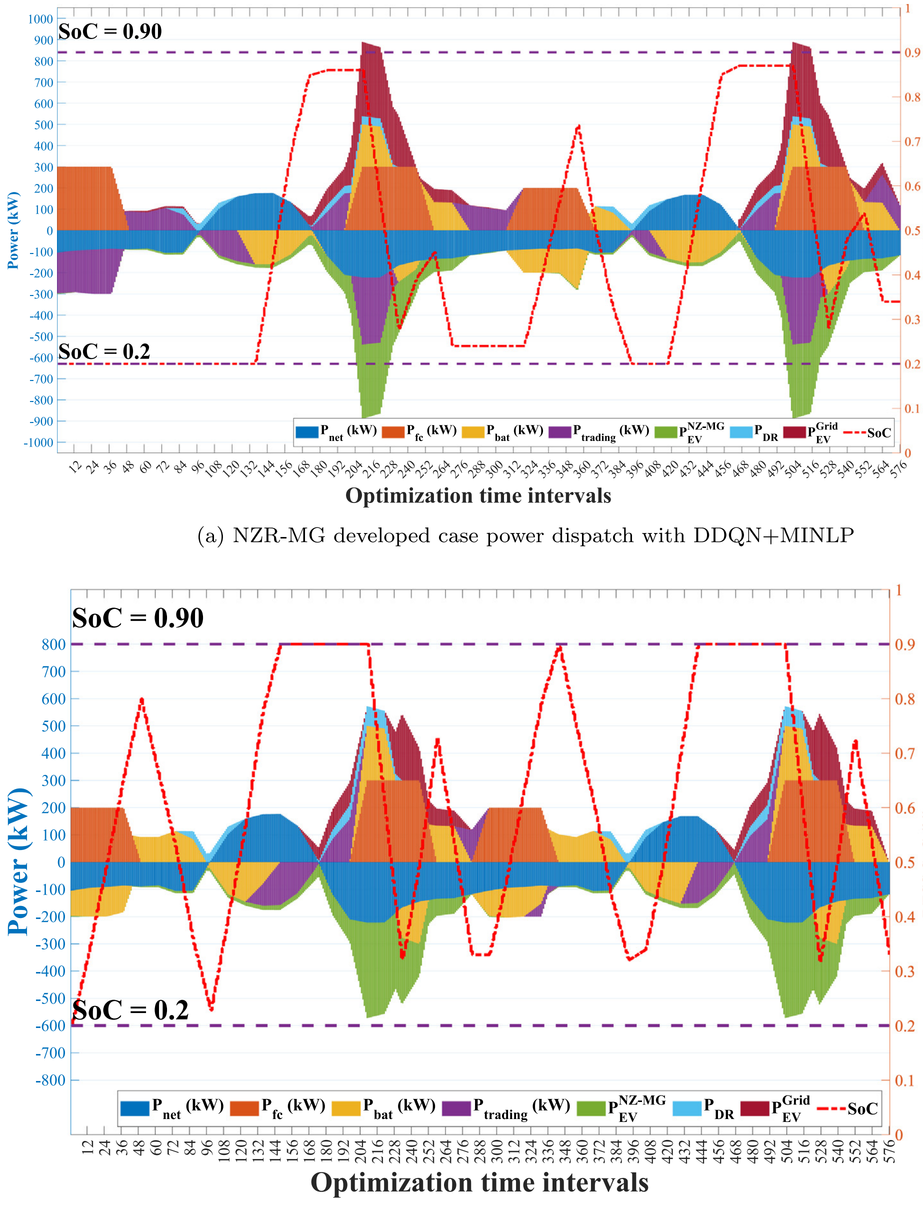
(b) NZR-MG base case hourly power dispatch with TL+DDQN+MINLP

Fig. 7. Comparison of NZR-MG base case hourly power dispatch with DDQN+MINLP and TL+DDQN+MINLP.

with two Intel (R) Cores (TM) i5-10400F CPU @2.90 GHz and 8 GB RAM. In this section, we investigated the performance of proposed algorithms for both base and developed cases of NZR-MG. Fig. 5 delineates a comparison of different algorithm's learning process average reward per episode over historical data from the training data set.

It can clearly be seen that TL+DDQN+MINLP converged in a shorter number of episodes comparing other algorithms. For base and developed cases, DDQN+MINLP stabilizes at 600 and 800 episodes, respectively, but adding TL to this algorithm drops the learning rate to 400 and 600. However, agent training convergence dramatically climbed to near 1000 for other methods. As is expected, DDPG and DQN learning processes have fluctuation. Unstable DDPG training results from hyper-parameter dependency and deterministic characteristics of the actor. It is observed from DDQN training that attaching a target network to DQN could overcome its unstable performance. The ϵ- greedy policy in action selection leads all agent performances to start from lower rewards. The exploitation approach guides agents to higher rewards with the increasing number of episodes. All algorithms in the base case converge faster than in the developed case because of its simpler state and action space.

Figs. 6 and 7 compare every 5-min power dispatch scheduling of the base case NZR-MG with proposed modified DDQNs for two sequential days in January from our test data set. Although Figs. 6(a) and 6(b) show both methods respect power balance in planning NZR-MG EMS, there is a difference in how algorithms retrain resource deployment balance. A general rule is that TL+DDQN+MINLP compensates for RESs absence by hiring internal resources, while DDQN+MINLP favors trading more power with the utility grid. As a result of this behavior, TL+DDQN+MINLP keeps the price of three-stage in the second stage,



(a) NZR-MG developed case power dispatch with DDQN+MINLP

(b) NZR-MG developed case power dispatch with TL+DDQN+MINLP

Fig. 8. Comparison of NZR-MG basecase hourly power dispatch with DDQN+MINLP and TL+DDQN+MINLP.

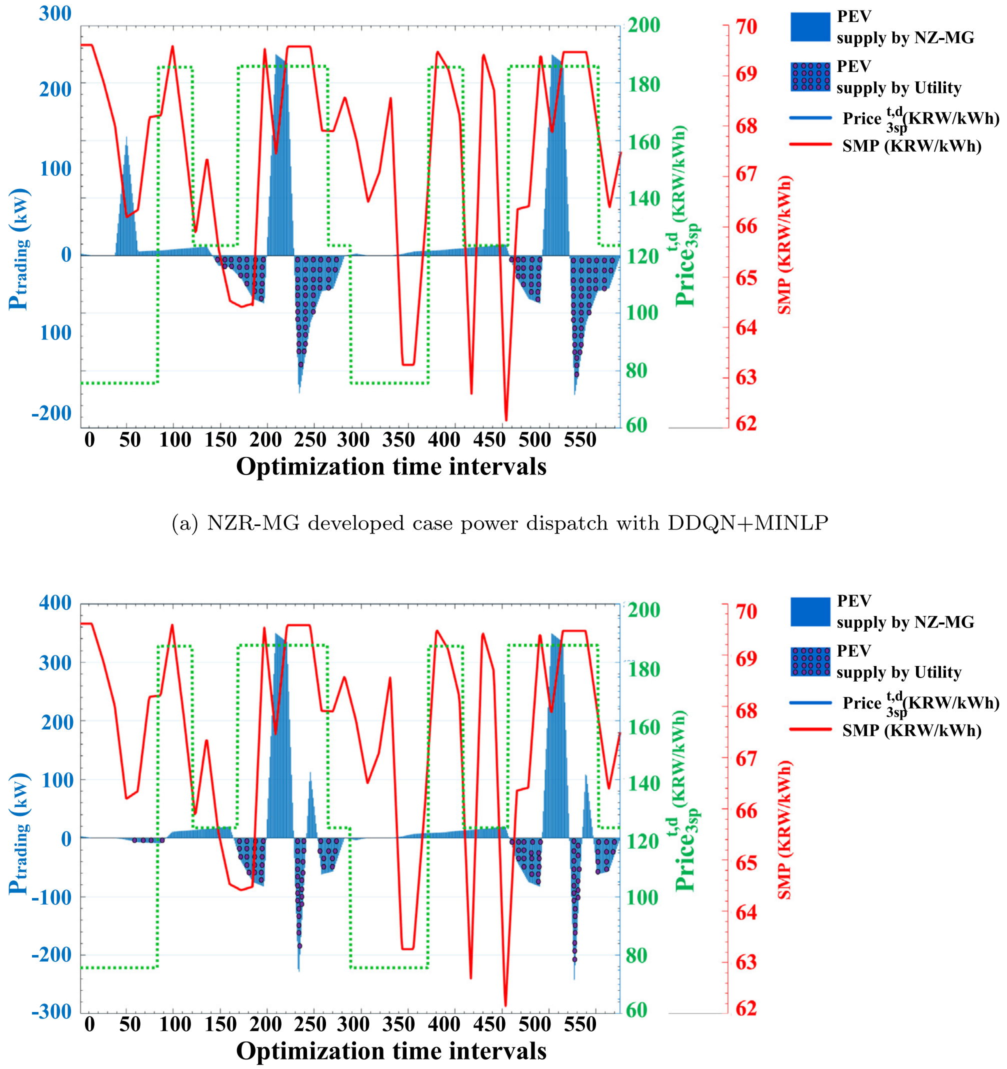
Table 1
NZR-MG annual cost(KRW/kWh) based on each optimization algorithm.

Algorithm	NZR-MG basecase	NZR-MG developed case	
	Operational cost	Operational cost	User anxiety cost
DQN	0.797×10^9	0.936×10^9	0.139×10^4
DDQN	0.763×10^9	0.871×10^9	0.156×10^3
DDPG	0.784×10^9	0.894×10^9	0.261×10^3
DDQN-MINLP	0.757×10^9	0.837×10^9	0.152×10^3
TL-DDQN-MINLP	0.743×10^9	0.822×10^9	0.123×10^3
MINLP solver (CPLEX)	0.675×10^9	0.797×10^9	0.05×10^3

according to Fig. 7(b). While in DDQN+MINLP, being greedy with acquiring profits from selling power in higher SMP rocketed the price for DDQN+MINLP solution to the third stage around 6 a.m. of the second day, as shown in Fig. 7(a).

NZR-MG developed case, with both proposed algorithms, has the same approach as the base case in EMS scheduling illustrated in Figs. 8 and 9. TL+MINLP+DDQN agent for developed case utilizes internal resources to fulfill energy shortages and supply EV contrasting the MINLP+DDQN agent, which falls into the trap of increasing income by selling power to the grid and passing the second stage of price.

The other significant TL efficacy appears in the demand response action of the agent. The sequential process of historical data feature extraction leads TL-based DDQN with more precision apply the limitation of the desired temperature. The demand response event occurs two times a day to decrease the peak–average ratio, according to Fig. 10. The room temperature in TL is precisely within the desired temperature in both understudies two sequential days following the second demand response event. As for DDQN+ MINLP, the amount of room temperature is higher than TL+MINLP+DDQN. This performance resulted in less



(a) NZR-MG developed case power dispatch with DDQN+MINLP

(b) NZR-MG developed case power dispatch with TL+DDQN+MINLP

Fig. 9. Comparison of NZR-MG basecase hourly power dispatch with DDQN+MINLP and TL+DDQN+MINLP.

contribution to decreasing demand during the second peak time of the day for the agent of DDQN+ MINLP.

The most important conclusion drawn from attaching TL to MINLP+DDQN is that although MINLP offers a more realistic reward function for DDQN, the TL with step-by-step retrieving features of historical data train agent hyper-parameters accurately.

We also formalized optimum policy for both cases of NZR-MG with MINLP and compared its results from the CPLEX solver with all hired DRL algorithms in case of annual costs and each stakeholder profit represented in Table 1 and Fig. 11. We considered the amount of peak–average ratio gained from the demand response scheduling of each algorithm solution to compare the amount of profit for the utility grid. It has been observed from Table 1 that TL+DDQN+MINLP has lower operational costs for both cases of NZR-MG as well as less user anxiety cost of demand response program implementation in the developed case. Unlike TL+DDQN+MINLP, DQN has the highest cost in base and developed scenarios, and DDPG often sits between these two extremes. Regarding the profit, TL+DDQN+MINLP also provides near the optimal solution annual profit for whole stakeholders of NZR-MG compared to the other methods, according to Fig. 11. Another notable point observed in Fig. 11 is that the system operator, which has a significant effect on increasing penetration of NZR-MG through investment, makes the highest annual profit from NZR-MG near to optimum point by utilization of our proposed algorithm.

In our study, we arranged EMS as a central unit that directly communicates with the utility company to dispatch resources. This hypothesis is based on the present Korean power system architecture. Despite the desire to maintain the Korean power system monopoly structure over the generation, transmission, and distribution sectors in the past, there are fresh insights into the power system business model to facilitate joining RESs, EVs, and demand response scheduling as flexible power resources in the framework of local electricity markets [35]. As a result, there is a surge in efforts to establish retail and local markets, resulting in emerging micro-grid communities such as the NZR-MGs community. In these communities, there is no doubt about preserving the privacy of stakeholders' electricity consumption and generation data by using brokers in information exchange with the utility company. Although our business model can be

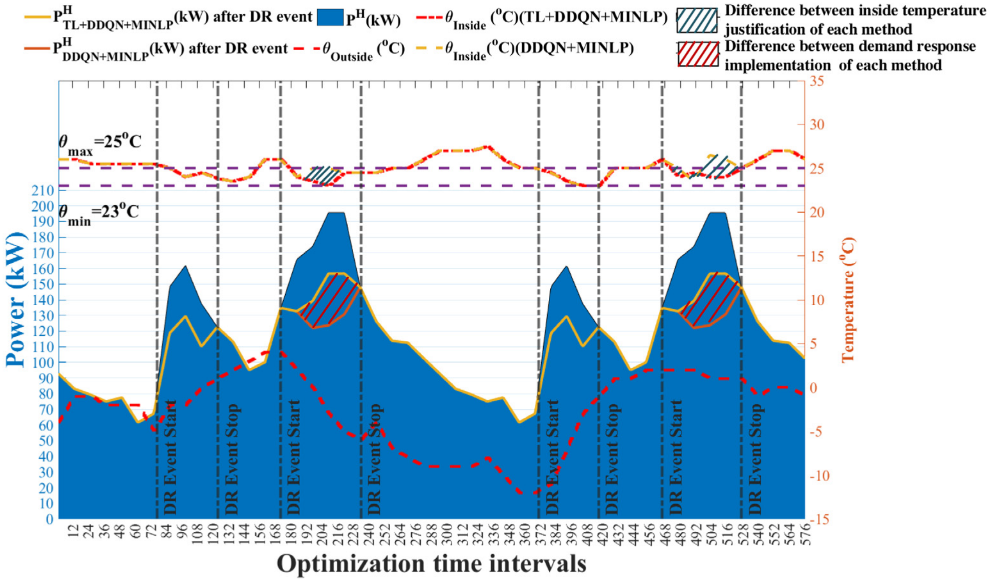


Fig. 10. Comparison of DDQN+MINLP and TL+DDQN+MINLP in demand response implementation.

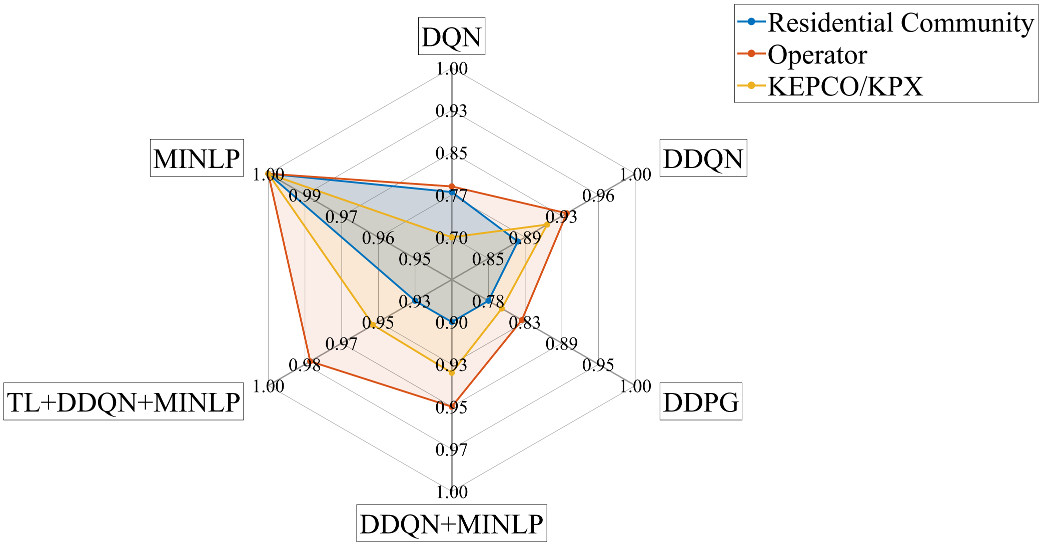


Fig. 11. Comparison of each NZR-MG stackholder profit provision based on utilized optimization algorithms.

extended to covering the brokers' interests based on evoking costs of services from stakeholders' relationships and arranging cost functions based on that, future work is required to establish the viability of privacy provision of stakeholders. The Multi-agent DRL techniques will meet privacy requirements, where our proposed method can be adapted to serve each micro-grid agent power dispatch locally.

The other open issue is the economic feasibility of our proposed technique in compensating investment expenses. We evaluated the viability of operational cost optimization, self-sufficiency, and excess generation management using the proposed algorithm. Therefore, further studies will be needed to prove the economic feasibility of the hired approach for NZR-MG business model arrangements in compensating investment costs.

7. Conclusion

This paper presents a profitable business model for the NZR-MG following Korean power system regulations and policies. A typical NZR-MG model in Korea is enhanced by the addition of EV charging stations and demand response plans, evidencing realistic recent power consumption trends. Our approach involves power dispatched scheduling of NZR-MG with a modified DDQN algorithm. The proposed algorithm could offer an online solution to maximize profit for whole stakeholders with the contribution of MINLP-based reward estimator and TL techniques. We compared our technique performance with a wide range of DQN-based algorithms. Precise hyper-parameter provision by breaking EMS task into several subtasks to hire TL and accurate reward approximation by MINLP solver employment guided the NZR-MG agent to the minimum difference with the optimum solution.

However, prospective alteration of the Korean power system monopoly structure to serve the competitive electricity market with RESs, EVs, and demand response scheduling will introduce brokers in NZR-MG and the utility company information exchange, which calls for the provision of privacy for stakeholders' data. To respect this privacy requirement, we will evaluate hiring multi-agent arrangements for our DRL approach in the future. Attaching investment expense compensation of NZR-MG to the

Table A.1
Acronyms.

Abbreviation	Description
A3C	Asynchronous advantages actor-critic
A2C	Advantage actor-critic
BESS	Battery energy storage systems
DDQN	Dual deep Q-learning (DDQN)
DDPG	Deep deterministic policy gradient
DER	Distributed energy resources
DG	Diesel generator
DNN	Deep neural network
DPG	Deterministic policy gradient
DQN	Deep Q-network
DRL	Deep reinforcement learning
DSO	Distribution system operator
EMS	Energy management system
ESS	Energy storage systems
EVs	Electric vehicles
GRU	Gated recurrent unit
KEPCO	Korean electric power company
KPX	Korean power exchange company
LSTM	Long short term memory neural networks
LV	Low voltage
MINLP	Mixed-integer nonlinear programming
MV	Medium-voltage
NZR-MG	Net-zero residential micro-grid
PCC	Point of common coupling
PPO	Proximal policy optimization
PVs	Photovoltaics
RES	Renewable energy sources
RL	Reinforcement learning
SAC	Soft-actor critic
SMP	System marginal price
SoC	state of the charge
TL	Transfer learning
TSO	Transmission system operator
V2G	Vehicle-to-grid
WT	Wind turbine

EMS problem cost function will be another issue that will need to be undertaken.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021R1F1A1063640).

Appendix A. Nomenclatures

See Tables A.1 and A.2.

Appendix B. Residential load and EV charging price in Korea

See Tables B.1 and B.2.

Appendix C. NZR-MG average monthly load profile estimation

Our NZR-MG is in Seoul, and it has 130 residential units located in three eight-story blocks, with an average of four occupants in each unit. We followed Korean household electricity consumption to provide real-world results. The primary electrical devices for each residential unit include a TV, refrigerator, Kimchi refrigerator, washing machine, rice cooker, and microwave, with power consumption taken from the Enertalk database [36]. Power usage is mainly affected by the heating and cooling system. Therefore, we considered the tri-generation system that facilitates heating, cooling, and hot water of buildings with the help of RESs and arranged average monthly load profiles [37]. Finally, the public electricity consumption of each building block is calculated by (C.1) from [38].

$$P_C(\Delta t) = P_P(0.148N_s + 0.092), \tag{C.1}$$

where P_C and P_P are the amounts of common and private electricity usage of each building block in each time slot Δt, respectively, and N_s denotes the number of floors. Fig. C.1 delineates estimated average monthly load profile of the NZR-MG during 2017 based on appliances, heating and cooling, and common usage.

Table A.2
Symbols.

Parameters		Description
Business model graph	i,j	Indices for the elements
	E	Set of relationships
	V	Set of elements
	ζ	Element's relationship
	K_a	Key action of the elements
	N	Number of elements
	$Profit(K_a^i\|\zeta_{i,j})$	Amount of each activity earning after distracting the cost
	$Cost_{deg}^-$	Cost of utilized technology degradation
	D^-	Customer power demand
	D^+	Customer power demand reduction du to utilizing technology
	π_{env}^+	Policy of the institution structure
	π_{DR}^+	Incentives for taking part in demand response
	$\pi_{Purchasing}^- \& \pi_{Selling}^+$	Policies for trading power with the utility grid
	u_a	Business model element's activating status
	E_G	Amount of elements energy generation
	E_C	Amount of elements energy consumption

(continued on next page)

Table A.2 (*continued*).

NZR-MG elements	PV	M	Number of PVs
		P_{PV}	PVs output power (kW)
		$P_{PV,min/max}$	PVs minimum/maximum output power (kW)
	Fuel cell	P_{FC}	The amount of fuel cell power output (kW)
		$T_{FC}^{up/down}$	Fuel cell's minimum up/down time
		$T_{FC}^{on/off}$	Fuel cell's duration of being on/off
		u_{FC}	A binary value to show the fuel cell is on or off
		$Cost_{Fuel}$	Cost of fuel cell's fuel (KRW)
		η_{FC}	Fuel cell efficiency (%)
	BESS	P_{bat}	Battery power charging or discharging (kW)
		$P_{bat,min/max}$	Battery minimum/maximum output power (kW)
		$SoC_{min/max}$	Battery state of the charge minimum/maximum
		E_b	Battery capacity (kWh)
		Δt	Time slot of the battery charging/discharging
		η_{bat}	Battery charging and discharging efficiency
		α	Battery degradation coefficient
	Residential load	P_C	Common electricity usage of each building block
		Δt	Time period in load consumption
		P_P	Private usage of building block units
		N_s	Number of building's floors
		$\theta_{in/out}$	Indoor/outdoor temperature
		$\theta_{min/max}$	Minimum/maximum desired Indoor temperature
		$\mathcal{P}_{DR}$	Participation probability for each demand response unit
		u_{ax}	Anxiety coefficient of consumers
		$P^{H\&C}$	Heating and cooling system power usage (kW)
		$P^{H\&C}_{min/max}$	Minimum/maximum heating and cooling system power usage (kW)
		$\mathcal{K}_1\&\mathcal{K}_2$	Coefficients to determine indoor temperature
		$\Delta P^{H\&C}$	H&C system power reduction to contribute in demand response (kW)
		$P_{DR,dev}$	Residential load reduction to contribute in demand response (kW)
	EV charging station	$\mathcal{P}(t, EV)$	the possibility of EV arriving home at time t
		δ_{tarr}	EV's arriving home standard deviation
		μ_{tarr}	EV's arriving home average value
		k	Number of charging piles
		$P_{n,rated}$	Rated power of charging piles
		ρ	Coefficient to determine EV charging power supplier
		$P_{CS,dev}$	Charging station power consumption (kW)
		$Price_{EV}^{t,d}$	EV charging power purchase rate (KRW/kWh)
	Utility grid	$Price_{EV}^{KEPCO}$	EV charging power purchase rate determined by utility grid (KRW/kWh)
		$Price_{3sp}^{t,d}$	Three-stage progressive rate energy price(KRW/kWh)
		$SMP^{t,d}$	The price of selling power to the utility grid (KRW/kWh)
		$P_{purchase}(.\|load)(t)$	The amount of power purchased from the utility grid to supply load
		$P_{purchase}(.\|EV)(t)$	The amount of purchased power from the utility grid to supply EVs
		$P_{sell}(t)$	The amount of sold power to the grid at each time t
DRL		$\mathcal{S}$	State space
		$\mathcal{A}$	Action space
		$\mathcal{T}$	Transition function
		$\mathcal{R}$	Reward function
		$\mathcal{A}_{cont/dics}$	Continuous/discrete action space
		$A_{purchase,dev\|EV}$	Supplying EV from the utility grid or NZR-MG resources
		$A_{DR,dev}$	The state of reducible load participation in the demand response
		$\mathcal{R}_{SoC}$	Reward function to keep SoC in the desired range
		$\mathcal{R}_{balance}$	Reward function to keep balance between power generation and consumption
		y_t	target value
		γ	Discounting factor
		ε	Probability of random action
		τ	Target smoothing factor
		θ/θ'	Critic/target critic networks weights
		μ/μ'	Policy/target policy networks weights
		Ub	Set of MINLP upper bounds
		Lb	Set of MINLP Lower bounds

Table B.1
EV charging price (KRW/kWh) based on KEPCO regulation.

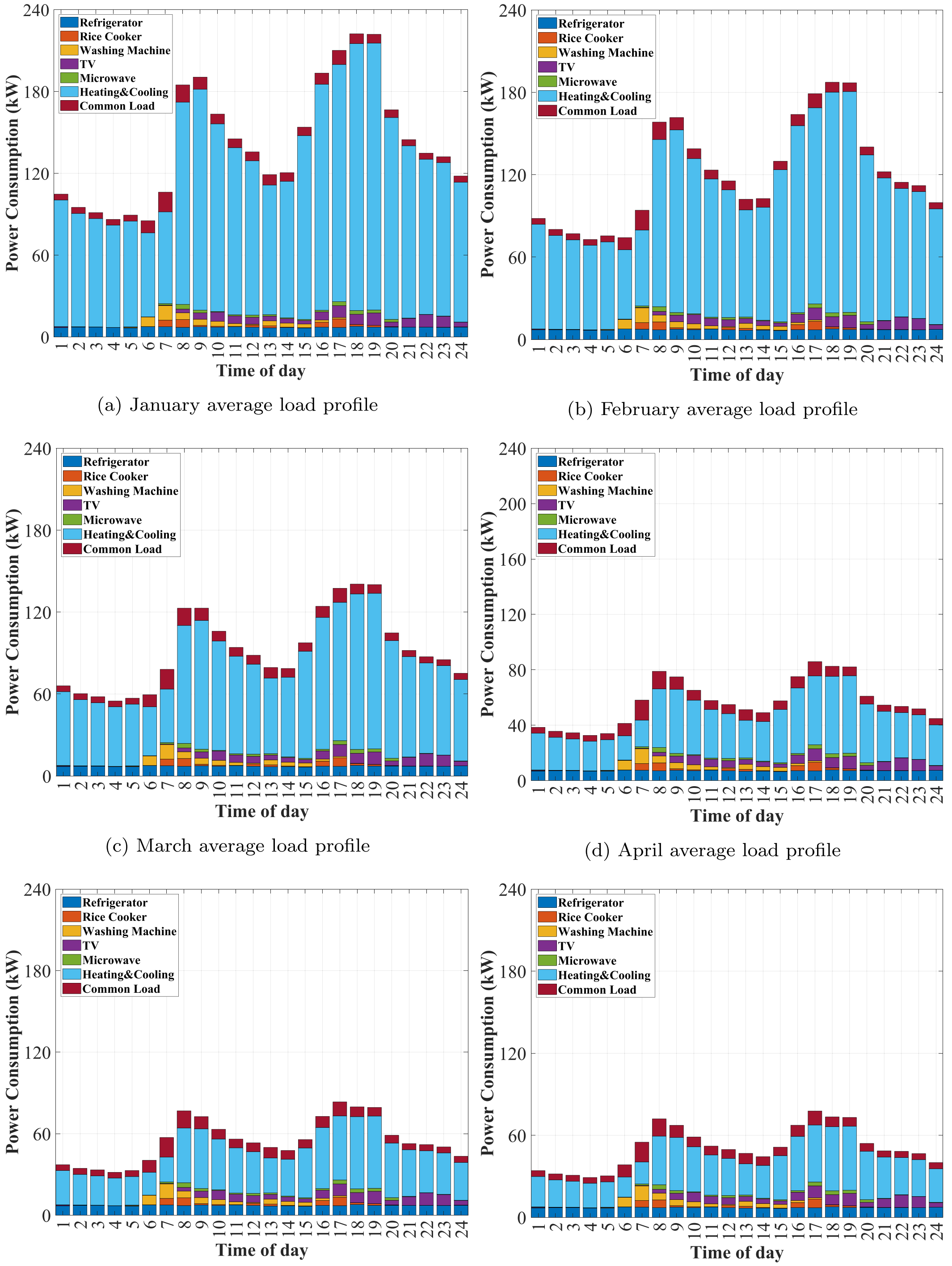
Time	Summer	Spring/Fall	Winter
Off-peak	52.6	53.7	75.7
Mid-peak	140.3	65.5	123.5
On-peak	227.5	70.4	185.8

Table B.2
Three-stage progressive price ($Price_{3sp}^{t,d}$) (KRW/kWh) based on KEPCO regulations.

Sum of energy consumption		Energy price
Summer (July 1∼Aug 31)	Other seasons	
1∼200 kWh	1∼300 kWh	73.3
201∼400 kWh	301∼450 kWh	142.3
401 kWh∼	451 kWh∼	210.6

Appendix D. NZR-MG elements technical specifications

See Table D.1.



(a) January average load profile

(b) February average load profile

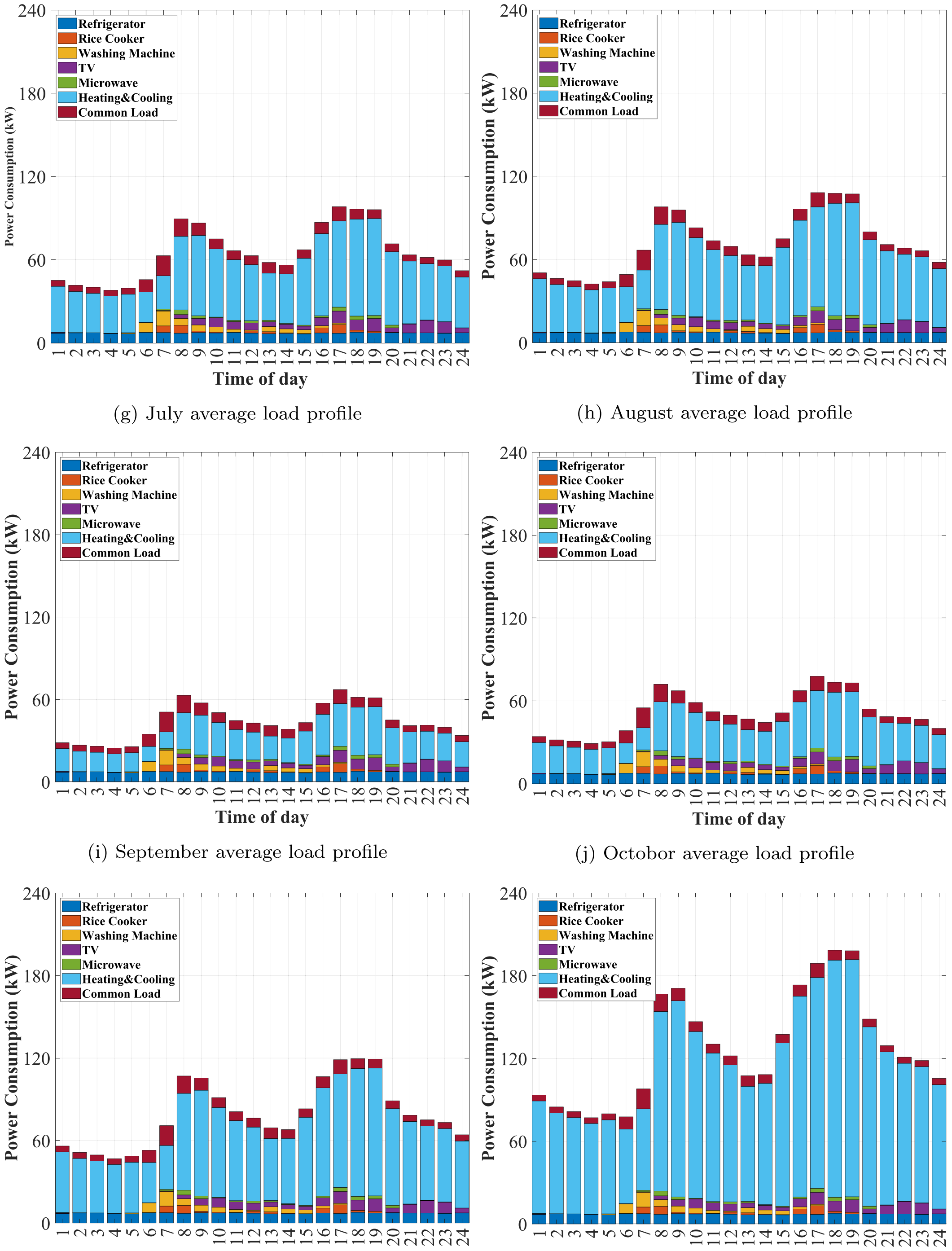
(c) March average load profile

(d) April average load profile

(e) May average load profile

(f) June average load profile

Fig. C.1. NZR-MG average load profile during the different months of 2017.



(g) July average load profile

(h) August average load profile

(i) September average load profile

(j) Octobor average load profile

(k) November average load profile

(l) December average load profile

Fig. C.1. (*continued*).

Table D.1
NZR-MG elements technical specification and constraints.

PV	$P_{PV,min}$ (kW)		0
	$P_{PV,max}$ (kW)		400
	$Cost_{PV}$ (KRW/kW)		0
Fuel cell	$P_{FC,min}$ (kW)		0
	$P_{FC,max}$ (kW)		300
	η_{FC} (%)		80
	$Cost_{Fuel}$ (KRW/kg)		8800
BESS	E_{bat} (kWh)		600
	$P_{bat,min}$ (kW)		−200
	$P_{bat,max}$ (kW)		200
	SoC_{min} (%)		20
	SoC_{max} (%)		90
	η_{bat} (%)		90
	α		0.9
Utility Grid	$P_{Purchase,max}$ (kW)		700
	$P_{Sell,max}$ (kW)		700
	Price of selling power to NZR-MG $Price_{3sp}^{t,d}$		acc. Table B.1
	Price of purchasing power from NZR-MG		$SMP^{t,d}$
Residential load	Load profile		acc. to Fig. C.1
	Demand response coefficients	θ_{min} (°C)	23
		θ_{max} (°C)	25
		u_x	100
		$\mathcal{K}_1$, $\mathcal{K}_2$	0.8, −0.02
EV	EVCS rated power		7 ∗ 50 kW
	EV arriving time	Pattern	acc. (22)
		δ_{tarr} (km)	38.8
		μ_{tarr} (km)	21.9
	Price of purchasing power		acc. (25)

References

[1] Tightiz L, Yang H, Bevrani H. An interoperable communication framework for grid frequency regulation support from microgrids. Sensors 2021;21(13):4555.

[2] Zhou B, Zou J, Yung Chung C, Wang H, Liu N, Voropai N, Xu D. Multi-microgrid energy management systems: Architecture, communication, and scheduling strategies. J Mod Power Syst Clean Energy 2021;9(3):463–76.

[3] Chen B, Wang J, Lu X, Chen C, Zhao S. Networked microgrids for grid resilience, robustness, and efficiency: A review. IEEE Trans Smart Grid 2021;12(1):18–32.

[4] Kim S. National competitive advantage and energy transitions in Korea and Taiwan. New Political Econ 2020;26(3):359–75.

[5] Huh T. Analyzing the configuration of knowledge transferof the Green Island projects in S. Korea. Korean J Policy Stud 2017;32(1):1–26.

[6] Hwang S, Lee Y, Sim J, Kim W, Lee H. Analysis on the stakeholders of microgrid businesses for the development of dissemination policies. In: CIRED workshop 2018. Ljubljana, Slovenia; 2018.

[7] Mocanu E, Mocanu D, Nguyen P, Liotta A, Webber M, Gibescu M, Slootweg JG. On-Line building energy optimization using deep reinforcement learning. IEEE Trans Smart Grid 2019;10(4):3698–708.

[8] Bui Y, Hussain A, Kim H. Double deep Q -learning-based distributed operation of battery energy storage system considering uncertainties. IEEE Trans Smart Grid 2020;11(1):457–69.

[9] Tightiz L, Yang H. Resilience microgrid as power system integrity protection scheme element with reinforcement learning based management. IEEE Access 2021;9:83963–75.

[10] Nakabi T, Toivanen P. Deep reinforcement learning for energy management in a microgrid with flexible demand. Sustain Energy Grids Netw 2021;25:100413.

[11] Shuai H, He H. Online scheduling of a residential microgrid via Monte-Carlo tree search and a learned model. IEEE Trans Smart Grid 2021;12(2):1073–87.

[12] Yoldas Y, Goren S, Onen A. Optimal control of microgrids with multi-stage mixed-integer nonlinear programming guided q-learning algorithm. J Mod Power Syst Clean Energy 2020;8(6):1151–9.

[13] Qin Z, Liu D, Hua H, Cao J. Privacy preserving load control of residential microgrid via deep reinforcement learning. IEEE Trans Smart Grid 2021;12(5):4079–89.

[14] Vatanparvar K, Al Faruque M. Design space exploration for the profitability of a rule-based aggregator business model within a residential microgrid. IEEE Trans Smart Grid 2015;6(3):1167–75.

[15] Hanna R, Disfan V, Kleissl J, Victor D. A new simulation model to develop and assess business cases for commercial microgrids. In: 2017 North American power symposium (NAPS). Morgantown, WV, USA; 2017.

[16] Qu M, Ding T, Huang L, Wu X. Toward a global green smart microgrid: An industrial park in China. IEEE Electrif Mag 2020;8(4):55–69.

[17] Lakshmi E S, Singh S, Padmanaban S, Leonowicz Z, Holm-Nielsen J. Prosumer energy management for optimal utilization of bid fulfillment with EV uncertainty modeling. IEEE Trans Ind Appl 2022;58(1):599–611.

[18] Sutton RS, Barto AG. Reinforcement learning, second edition: An introduction. MIT Press; 2018.

[19] Reforming Korea's electricity market for net zero. Technical Report 127, International Energy Agency (IEA)/ Korea Energy Economics Institute (KEEI); 2021, URL https://www.iea.org. Accessed=2022-08-17.

[20] Vanadzina E, Mendes G, Honkapuro S, Pinomaa A, Melkas H. Business models for community micro-grids. In: 16th international conference on the European energy market (EEM). Ljubljana, Slovenia; 2018.

[21] Choi S, Min S. Optimal scheduling and operation of the ESS for prosumer market environment in grid-connected industrial complex. IEEE Trans Ind Appl 2018;54(3):1949–57.

[22] Baek M, Shin B. Hybrid operation strategy for demand response resources and energy storage system. J Electr Eng Technol 2021;17(1):25–37.

[23] Ryu J, Kim J. Non-Cooperative indirect energy trading with energy storage systems for mitigation of demand response participation uncertainty. Energies 2020;13(4):883.

[24] Lee S, Choi D. Reinforcement learning-based energy management of smart home with Rooftop Solar Photovoltaic System, energy storage system, and home appliances. Sensors 2019;19(18):3937.

[25] Arias M, Bae S. Electric vehicle charging demand forecasting model based on big data technologies. Appl Energy 2016;183:327–39.

[26] Kim J, Kim C. Demand power with EV charging schemes considering actual data. J Int Counc Electr Eng 2016;6(1):235–41.

[27] Hussain A, Bui V, Kim H. Optimal sizing of battery energy storage system in a fast EV charging station considering power outages. IEEE Trans Transp Electrif 2020;6(2):453–63.

[28] Sachan S, Deb S, Singh S. Different charging infrastructures along with smart charging strategies for electric vehicles. Sustainable Cities Soc 2020;60:102238.

[29] Korean Electric Power Company. Korea electric rate table. 2022, URL https://home.kepco.co.kr/kepco/EN. Accessed=2022-04-01.

[30] Korean Power Exchange Company. Korean electric power statistics information system system marginal price. 2022, URL http://epsis.kpx.or.kr/epsisnew/selectEkmaSmpSmpGrid.do?menuId=040201&locale=eng. Accessed=2022-04-01.

[31] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. Playing atari with deep reinforcement learning. 2013, arXiv preprint arXiv:1312.5602.

[32] Van H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the AAAI conference on artificial intelligence. Phoenix, Arizona, USA; 2016.

[33] Lillicrap T, Hunt J, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. 2015, arXiv:1509.02971.

[34] Open energy system databases § renewables.ninja. 2022, URL http://www.renewables.ninja. Accessed=2022-04-01.

[35] Hyun S, Kim C, Lee B. KEPCO's movement on distribution sector regarding renewable energy transition of distribution network in Korea. KEPCO J Electr Power Energy 2021;7(1):93–9.

[36] Shin C, Lee E, Han J, Yim J, Rhee W, Lee H. The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea. Sci Data 2019;6(1).

[37] Bae S, Nam Y, Cunha I. Economic solution of the tri-generation system using photovoltaic-thermal and ground source heat pump for zero energy building (ZEB) realization. Energies 2019;12(17):3304.

[38] Cheong C, Park B, Ryu S. A modified energy evaluation tool for residential complexes in South Korea to reflect total electricity consumption. J Asian Archit Build Eng 2017;16(1):215–22.