Titre: Title:	Multi-agent deep reinforcement learning with online and fair optimal dispatch of EV aggregators
Auteurs: Authors:	Arian Shah Kamrani, Anoosh Dini, Hanane Dagdougui, & Keyhan Sheshyekani
Date:	2025
Type:	Article de revue / Article
Référence: Citation:	Kamrani, A. S., Dini, A., Dagdougui, H., & Sheshyekani, K. (2025). Multi-agent deep reinforcement learning with online and fair optimal dispatch of EV aggregators. Machine Learning with Applications, 19, 100620 (12 pages). https://doi.org/10.1016/j.mlwa.2025.100620

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/61958/
Version:	Version officielle de l'éditeur / Published version Révisé par les pairs / Refereed
Conditions d'utilisation: Terms of Use:	Creative Commons Attribution-Utilisation non commerciale-Pas d'oeuvre dérivée 4.0 International / Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND)

Titre de la revue: Journal Title:	Machine Learning with Applications (vol. 19)
Maison d'édition: Publisher:	Elsevier
URL officiel: Official URL:	https://doi.org/10.1016/j.mlwa.2025.100620
Mention légale: Legal notice:	©2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by- nc-nd/4.0/).

Machine Learning with Applications 19 (2025) 100620

Contents lists available at ScienceDirect

Machine Learning with Applications

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

Multi-agent deep reinforcement learning with online and fair optimal dispatch of EV aggregators

Arian Shah Kamrani [a,b,*], Anoosh Dini [c], Hanane Dagdougui [a,b], Keyhan Sheshyekani [c]

[a] *Department of Mathematics and Industrial Engineering, Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, QC H3T 1J4, Canada*
[b] *GERAD Research Center, 3000 Chemin de Polytechnique, Montréal, QC H3T 2A7, Canada*
[c] *Department of Electrical Engineering, Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, QC H3T 1J4, Canada*

ARTICLE INFO

Keywords:
Electric vehicle (EV)
Reinforcement learning (RL)
Energy management
Jain's index
Distribution system operator (DSO)

ABSTRACT

The growing popularity of electric vehicles (EVs) and the unpredictable behavior of EV owners have attracted attention to real-time coordination of EVs charging management. This paper presents a hierarchical structure for charging management of EVs by integrating fairness and efficiency concepts within the operations of the distribution system operator (DSO) while utilizing a multi-agent deep reinforcement learning (MADRL) framework to tackle the complexities of energy purchasing and distribution among EV aggregators (EVAs). At the upper level, DSO calculates the maximum allowable power for each EVA based on power flow constraints to ensure grid safety. Then, it finds the optimal efficiency-Jain tradeoff (EJT) point, where it sells the highest energy amount while ensuring equitable energy distribution. At the lower level, initially, each EVA acts as an agent employing a double deep Q-network (DDQN) with adaptive learning rates and prioritized experience replay to determine optimal energy purchases from the DSO. Then, the real-time smart dispatch (RSD) controller prioritizes EVs for energy dispatch based on relevant EVs information. Findings indicate the proposed enhanced DDQN outperforms deep deterministic policy gradient (DDPG) and proximal policy optimization (PPO) in cumulative rewards and convergence speed. Finally, the framework's performance is evaluated against uncontrolled charging and the first come first serve (FCFS) scenario using the 118-bus distribution system, demonstrating superior performance in maintaining safe operation of the grid while reducing charging costs for EVAs. Additionally, the framework's integration with renewable energy sources (RESs), such as photovoltaic (PV), demonstrates its potential to enhance grid reliability.

1. Introduction

1.1. Motivation

In recent years, electric vehicles (EVs) have attracted attention for their environmentally friendly nature and higher efficiency compared to fossil-fueled vehicles (Bai et al., 2024). The growing number of EVs, however, may pose challenges such as voltage fluctuations, transformer overloading, and power outages (Bao, Hu, & Mujeeb, 2024). Additional challenges in the charging process of EVs arise from uncertainties associated with EV users' behavior, non-EV load demand, renewable energy sources (RESs), and fluctuations in electricity prices (Madahi, Kamrani, & Nafisi, 2022). On the other hand, EVs provide flexibility, allowing them to participate in energy markets (Qi, Liu, Lu, Yu, & Degner, 2023), demand response programs (Jin, Zhou, Lu, & Song, 2022), and ancillary services (Kiani, Sheshyekani, & Dagdougui, 2023) facilitated by aggregators. Consequently, the efficient management of EV charging is of significant importance for both DSO and aggregators.

EVs online charging management has consistently presented challenges for DSO and aggregators due to the uncertain nature of this problem. Traditional EV charging strategies, based on optimization, require substantial computational resources, so their performance is heavily reliant on prior knowledge of the system (Zhang, Rao, Liu, Zhang, & Zhou, 2023). In contrast to the traditional approaches, reinforcement learning (RL) enables learning optimal control strategies through the interaction between agents and the environment, without the need for an accurate system model and uncertainties (Ding et al., 2020). Moreover, unlike offline and prediction-based optimization methods which are computationally intensive, learning-based strategies offer a balance between online implementation and achieving optimal solutions (Arwa

[*] Corresponding author at: Department of Mathematics and Industrial Engineering, Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, QC H3T 1J4, Canada.

E-mail addresses: arian.shah-kamrani@polymtl.ca (A.S. Kamrani), anoosh.dini@polymtl.ca (A. Dini), hanane.dagdougui@polymtl.ca (H. Dagdougui), keyhan.sheshyekani@polymtl.ca (K. Sheshyekani).

https://doi.org/10.1016/j.mlwa.2025.100620
Received 21 November 2024; Received in revised form 23 December 2024; Accepted 3 January 2025
Available online 9 January 2025

& Folly, 2020). Hence, RL is proficient at efficiently solving sequential decision problems in complex and uncertain environments, including EVs real-time charging optimization.

1.2. Related works

In previous years, EV fleet charging management utilizing RL algorithms has increased attention (Abdullah, Gastli, & Ben-Brahim, 2021). Cao, Wang, Li, and Zhang (2022) proposed an RL-based smart charging algorithm to reduce the charging cost while considering peak load shaving under uncertainties related to EV owners. Moreover, an enhanced customized actor–critic learning algorithm is introduced to reduce the state dimension and thus improve the computational efficiency. Zhang, Yang, An, Li, and Wu (2023) employed a multiagent deep deterministic policy gradient approach to acquire the optimal energy purchasing strategy for charging stations. Additionally, an online heuristic dispatching scheme is proposed to formulate an energy distribution strategy among EVs. Zhang, Liu, Wu, Tang, and Fan (2021) initially introduced the EV charging scheduling problem and the NP-hardness of the problem is demonstrated. Subsequently, the scheduling problem of EV charging is formalized as a Markov decision process (MDP), and deep RL algorithms are suggested for its resolution. The algorithms proposed aim to minimize the total charging time of EVs and achieve a maximal reduction in the origin–destination distance. While this paragraph discusses existing research in the domain of RL-based EV fleet charging, notable shortcomings persist in the field. The mentioned works can be broadly categorized into two groups: the first focuses on centralized management with a single controller entity utilizing single-agent RL, while the second employs multi-agent reinforcement learning (MARL) for decentralized optimization, treating each EV as an agent. Notably, to the best of the authors' knowledge, only few studies consider the EV aggregator (EVA) as an agent in a MARL environment for maximizing its benefit by managing the purchasing and distribution strategy of energy among EVs simultaneously. Additionally, in the domain of RL-based charging management of EVs, there is a lack of work addressing the acceleration of RL algorithm convergence speed for real-time purposes.

As modern society continues to advance, the demand for electricity is steadily rising, leading to an increase in the scale and complexity of power systems (Wang, Chen, Zhou, Liu, & Peng, 2024). Allocating resources among users is a common challenge in any distributed system. In power distribution systems, a *fair* allocation ensures each consumer receives an equitable share of power. However, an *efficient* allocation is achieved when the DSO sells the maximum allowable power to each consumer. While DSO's goal is to maximize efficiency, consumers are more inclined to maximize their benefits, often resulting in conflicts of interest (Sediq, Gohary, Schoenen, & Yanikomeroglu, 2013). Fairness can be incorporated into any resource allocation problem, regardless of its application. However, it is often overlooked in power system studies, with only a few research investigating it. Hupez, Toubeau, De Grève, and Vallée (2021) developed a structure to exchange energy within a low voltage community whose goal is to minimize the overall community cost. The structure exploits the resilience offered by surplus storage and generation capacity, and then it uses the Nash equilibrium to fairly share the overall cost among members. However, the approach is purely economic-centered, failing to consider equitable access to energy resources among members. On the other hand, Hussain and Musilek (2022) focuses on fair energy allocation during power contingencies. It utilizes Jain's fairness index (JFI) to evaluate and enhance fairness in energy allocation among EVs during outages. However, the study focuses on just one charging station and uses simulations that update every hour, which is not practical for real-time use. Other studies on fairness within power systems scope focus on PV curtailment during off-peaks to avoid reverse power flow and overvoltage issues (Poudel, Mukherjee, Sadnan, & Reiman, 2023). The mentioned research considers fairness from the perspective of load balancing, not from the standpoint of equitable resource allocation.

Unlike fairness, the literature on hierarchical frameworks on EV charging scheduling is rich. At the EV level, the main goal of these frameworks is to fulfill the charging demands of EV users. However, at the DSO and aggregator levels, these frameworks aim to tackle goals, such as peak shaving (Wu, Radhakrishnan, & Huang, 2019), reducing demand charges (Saner, Trivedi, & Srinivasan, 2022), minimizing EV charging costs (Kiani, Sheshyekani, & Dagdougui, 2024), and privacy preservation (Amini, McNamara, Weng, Karabasoglu, & Xu, 2018). In this context, (Wu et al., 2019) proposed a two-level hierarchical framework to involve EVs in peak shaving while meeting users' demands.

However, the paper primarily focuses on demand management and grid services facilitation without considering the power flow constraints to ensure the safe operation of the grid. Additionally, a distributed model predictive control-based strategy for multiple EV charging stations is proposed by Zheng, Song, Hill, and Meng (2019). While the proposed strategy is online and capable of ensuring the grid's safe operation, it depends on the charging station operator having access to the grid's parameters. However, this may not always be feasible due to concerns over data privacy and the varying economic interests of different charging stations.

Furthermore, Saner et al. (2022) formulated the EV charging scheduling problem as a multi-agent based optimization in a distributed and privacy-preserved manner. Although the grid's security constraints are satisfied in the proposed scheme, it initially sets a fixed charging schedule for EVs upon their arrival without reassessing the charging plans based on subsequent changes in the grid's load.

1.3. Contributions

To address the previously mentioned issues, an online safe multi-agent deep reinforcement learning (MADRL) framework for energy purchasing and distribution of EVAs has been introduced. This framework ensures data privacy for all entities while simultaneously considering their respective benefits. Moreover, in the proposed MADRL framework, a double deep Q-network (DDQN) featuring adaptive learning rates and a prioritized experience replay is employed to enhance the convergence of the agent towards optimality and make it particularly effective for online applications. Last but not least, at the level of DSO, energy dispatch among EVAs has been done based on optimal efficiency-Jain tradeoff (EJT) and constraint of the distribution network to ensure fairness and reliability of the framework.

Overall, the key contributions of this study can be summarized as follows:

- Presenting an online safe MADRL framework aimed at minimizing the purchasing and distribution costs of EVAs while accounting for distribution network constraints, which incorporates a DDQN algorithm featuring adaptive learning rates and prioritized experience replay for accelerated convergence compared to traditional RL techniques.
- Ensuring the safe operation of the grid by calculating the maximum allowable power based on real-time grid conditions, and leveraging these calculations to determine the optimal EJT point. This approach guarantees fair power allocation among EVAs while enabling the DSO to sell the maximum amount of power at each time step.
- Preserving the privacy of EV owners, EVAs, and DSO, eliminating the need for them to share their respective data in the proposed hierarchical structure. Furthermore, the structure prioritizes the mutual benefits of all stakeholders involved simultaneously.

The remainder of this paper is structured as follows: Section 2 provides an overview of the proposed framework. In Section 3, we provide a detailed explanation of the proposed MADRL model, including the method for determining the maximum allowable power and the integration of fairness into our work. Section 4 discusses the case study, followed by an analysis of simulation experiments in Section 5. Finally, conclusion and future work are summarized in Section 6.

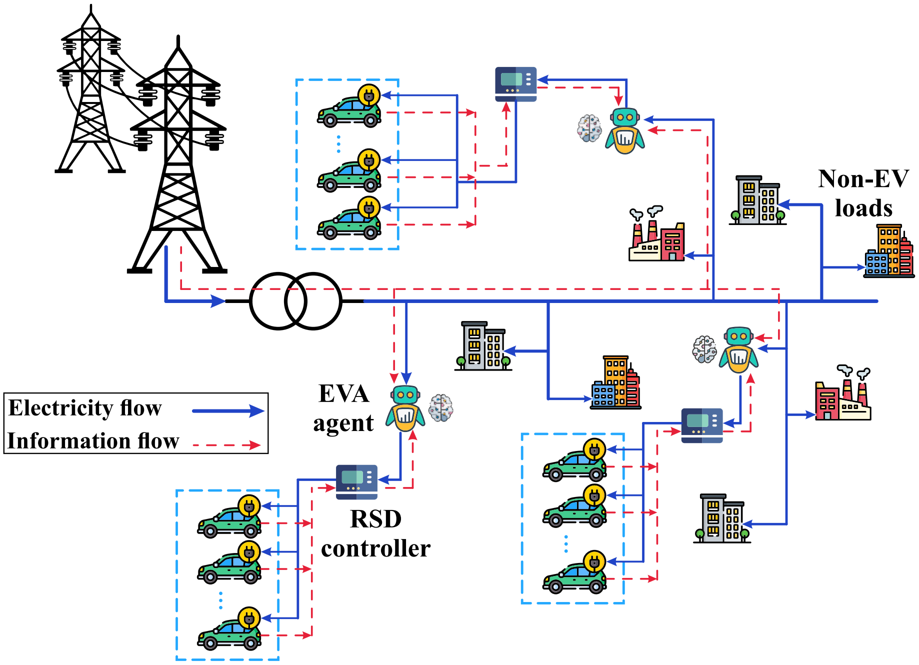


Fig. 1. Proposed framework.

2. Proposed framework

2.1. Developed model

The top-down approach of our proposed hierarchical framework is shown in Fig. 1. At the upper level, the responsibility of assuring the reliability of the distribution network and fair energy allocation to EVAs is held by the DSO. In this framework, the DSO initiates the calculation of the maximum allowable power allocated (i.e., safe margin) to each EVA at each time step. This computation is based on non-EV loads, the location of each EVA on the distribution network, and load flow constraints. Moreover, at the level of DSO, optimal EJT strategy is employed to assure fairness in energy allocation among EVAs. Ultimately, a signal set point indicating the maximum power available for purchase is transmitted from the DSO to each EVA at every time step. At the lower level, which is comprised of two sub-levels, first, EVAs act as agents within the proposed MADRL structure, determining the optimal amount of energy to purchase from DSO at each time step. They make decisions based on data provided by the DSO and RSD controller, which includes electricity price, the average state of charge (SoC) of their associated EVs, and the number of EVs involved. The amount of energy purchased should be carefully calibrated to meet the needs of EVs without being too low, as well as avoiding excess that could result in wastage. Afterward, the RSD controller conducts EV prioritization for energy dispatch at each time step by sorting the EVs based on their departure time, SoC, and maximum charging rate. This process involves considering the energy acquired by EVAs and subsequently dispatching it among the EVs. In this framework, when it comes to preserving the privacy of stakeholders, the DSO is not required to share any sensitive data regarding the distribution network topology with the EVAs. Furthermore, EVAs avoid sharing consumer data with both each other and the DSO. The only data that each EVA receives is generalized cumulative information about its associated EVs from RSD controller. From the perspective of stakeholders' benefits, the DSO ensures distribution system reliability and scalability, EVAs receive a fair energy allocation to maximize their profits, and EV owners obtain an energy dispatch service from RSD that considers the lifespan of EV batteries and the desired final SoC.

2.2. Application of MADRL

Traditional optimization methods struggle to handle the dynamic and multi-dimensional nature of real-time decision-making, such as the charging management of EVs. These methods often fail when faced with a high degree of uncertainty related to EV owners' behavior and fluctuating real-time electricity prices (Paudel & Das, 2023). However, MADRL is highly effective in these environments because it continuously learns and adapts to changes, handles complex interactions between agents, and is well-suited for managing real-time energy distribution and EV charging challenges (Yuan, Forhad, Bansal, Sidorova, & Albert, 2024). Moreover, MADRL supports scalability by enabling a large number of EVAs to operate and learn in parallel without the need for centralized control, thereby reducing computational overhead. Finally, in terms of privacy, since each EVA (agent) only needs generalized cumulative information and does not require access to detailed consumer data and the entire network topology, privacy is preserved for all stakeholders.

3. Problem formulation

3.1. Safe margin power

The safe margin power is a parameter obtained by DSO which guarantees the safe operation of the grid. This parameter indicates the maximum power that each EVA is allowed to draw at each time step. Inspired by Saner et al. (2022), we propose an optimization problem to calculate this value. In this regard, let N denotes the total number of buses in the grid. For each bus n, if an EVA is present, its associated safe margin power is represented as $P_n^{agg}(t)$; otherwise, the value is set to 0. Therefore, the vector of safe margin power at time step t, $\mathbf{P}^{SM}(t)$, can be defined as a vector of length N, where each element corresponds to a bus in the grid. This vector will be obtained such that the sum of its elements is maximized. DSO calculates $\mathbf{P}^{SM}(t)|_{t\in T}$ online and dispatches equitably adjusted values of the vector to their respective EVAs. To achieve the first goal, DSO solves the following optimization problem at each time step:

$$\mathbf{P}^{SM}(t) = \arg \max_{\substack{V_n,\varphi_n,\\(n,b)\in\mathcal{B}}} \left\{ \sum_{n=1}^{N} \mathbf{1}_{\{n\in\mathcal{A}\}} \cdot P_n^{agg}(t) \right\} \tag{1a}$$

$$\text{s.t.}\quad P_{nb} = y_{nb}\left(V_n^2 \cos\left(\phi_{nb}\right) - V_n V_b \cos\left(\varphi_n - \varphi_b - \phi_{nb}\right)\right), \tag{1b}$$

$$Q_{nb} = y_{nb}\left(V_n^2 \sin(\phi_{nb}) - V_n V_b \sin(\varphi_n - \varphi_b - \phi_{nb})\right), \tag{1c}$$

$$\sum_{b:\, y_{nb}\neq 0} P_{nb} = \frac{P_n^g(t) - \left(P_n^d(t) + P_n^{agg}(t)\right)}{S_{base}}, \tag{1d}$$

$$\sum_{b:\, y_{nb}\neq 0} Q_{nb} = \frac{Q_n^g(t) - Q_n^d(t)}{S_{base}}, \tag{1e}$$

$$\underline{V_n} \le V_n \le \overline{V_n}, \tag{1f}$$

$$\underline{\varphi_n} \le \varphi_n \le \overline{\varphi_n}, \tag{1g}$$

$$\sqrt{P_{nb}^2 + Q_{nb}^2} \le \frac{\overline{S_{nb}}}{S_{base}}, \tag{1h}$$

where $\mathcal{A}$ and $\mathcal{B}$ represent the set of buses with EVAs and all the buses in the distribution system, respectively. Accordingly, the pair $(n, b) \in \mathcal{B}$ specifies the connections between buses, where n and b are indices of the connected buses. Additionally, $\mathbf{1}_{\{n\in\mathcal{A}\}}$ is an indicator function that equals 1 if bus n has an EVA (i.e., $n \in \mathcal{A}$), and 0 otherwise. This function ensures that $P_n^{agg}(t)$ contributes to the sum only if an EVA is present at bus n.

Furthermore, $V_n\angle\varphi_n$ denotes the per-unit voltage at bus n, and $y_{nb}\angle\phi_{nb}$ indicates the series admittance between buses n and b. The active and reactive power flows on the branch from bus n to bus b are represented by P_{nb} and Q_{nb}, respectively. The problem also considers active and reactive power generated ($P_n^g(t)$ and $Q_n^g(t)$) and consumed ($P_n^d(t)$ and $Q_n^d(t)$) at each bus n, along with the power consumption of EVAs ($P_n^{agg}(t)$) where applicable. Constraints (1f) to (1h) indicate voltage magnitude, voltage angle, and apparent power flow limits, respectively, which are operational constraints and ensure the system's safety and reliability. In this context, the overline and underline symbols indicate the upper and lower limits for parameters, respectively. Moreover, S_{base} represents the base power used for per-unit calculations.

As outlined in (1), the safe margin power indicates the maximum allowable power that can be drawn from each bus while respecting the grid's operational constraints. This question then arises: how much of this power should be allocated to the EVA connected to that bus? Allocating the entire amount to the corresponding EVA is one strategy. Nevertheless, this might not be equitable for other EVAs and loads. The fairness issue arises because an EVA's location in the grid and the loads on that branch mainly determine the safe margin value, making it unlikely for all EVAs to have equal safe margin values. Furthermore, assigning the entire safe margin to an EVA allows it to increase its power to this limit, possibly leading to line congestion or voltage drop. This is particularly problematic if other loads on the same branch also wish to increase their consumption, leading the DSO to restrict such increases to ensure grid stability. In the next section, we present a strategy to achieve the optimal EJT point that takes into account the DSO benefit and ensures fair assignment of the maximum power to each EVA.

3.2. Fair resource allocation

To effectively meet the demands of EV owners while maintaining the safe operation of the grid, it is essential to allocate a fair maximum power limit for each EVA. The fairness indicator used in this study (i.e., JFI) is originally used in communication networks but also applies to other fields (Jain, Chiu, Hawe, et al., 1984).

Let $\mathbf{P}$ be a vector whose elements indicate the benefits, i.e., the allocated power, determined by DSO for I EVAs. Consider $\mathbf{P} \in \mathcal{S} \subseteq \mathbb{R}_+^I$, where $\mathcal{S}$ contains all the possible benefit vectors bounded by the safety margins. The space $\mathbb{R}_+^I$ contains all vectors with dimensions I that have non-negative components. Each element of $\mathbf{P}$ (i.e., P_i) corresponds to the allocated power of the ith EVA.

Definition 1 (*JFI (Jain et al., 1984) and Efficiency*)**.** For $\mathbf{P} \in \mathbb{R}_+^I$, Jain's fairness index $J : \mathbb{R}_+^I \to \mathbb{I}_+$ and efficiency $\psi : \mathbb{R}_+^I \to \mathbb{R}_+$ are respectively obtained by

$$J(\mathbf{P}) = \left(\sum_{i=1}^{I} P_i\right)^2 \Bigg/ I \sum_{i=1}^{I} P_i^2, \tag{2}$$

$$\psi(\mathbf{P}) = \sum_{i=1}^{I} P_i, \tag{3}$$

where I represents the total number of EVAs. The fairness of this allocation is quantified by $J(\mathbf{P})$. It is a continuous function with a range from $\frac{1}{I}$ to 1. A value of $J(\cdot) = \frac{1}{I}$ corresponds to least fair allocation, where only one EVA receives a non-zero values. Conversely, $J(\cdot) = 1$ indicates the most fair distribution, with every EVA receiving an identical value as the allocated power.

There is often a trade-off between resource efficiency and JFI (Sediq et al., 2013). Our goal is to find the optimal balance between the two, maximizing efficiency while keeping JFI close to 1.

Definition 2 (*Optimal EJT (Sediq et al., 2013)*)**.** Let $\mathcal{S}$ be a set of vectors, and $\mathbf{P}^*$ be an element of $\mathcal{S}$. The element $\mathbf{P}^*$ is considered to be the optimal EJT if no $\mathbf{P} \neq \mathbf{P}^*$, satisfies either: (1) $\psi(\mathbf{P}) > \psi(\mathbf{P}^*)$, and at the same time, $J(\mathbf{P}) \ge J(\mathbf{P}^*)$, or (2) $\psi(\mathbf{P}) \ge \psi(\mathbf{P}^*)$, and at the same time, $J(\mathbf{P}) > J(\mathbf{P}^*)$.

The concept described in the above definition is similar to the situation in multi-objective optimization problems. Having both efficiency and JFI as objectives, an optimal point is reached when any attempt to improve one objective worsens the other. A technique to obtain the optimal EJT is presented by Sediq et al. (2013). In order to effectively represent the proposed technique, we first need to introduce some key terms. Let τ indicates the minimum efficiency. Additionally, we can define the set of benefit vectors that meet two criteria: they exceed the value of τ, while at the same time having the highest possible JFI. We refer to this set as:

$$\mathcal{P}_\tau \triangleq \left\{ \mathbf{P} | \mathbf{P} = \arg \max_{\psi(\mathbf{P})\ge\tau,\ \mathbf{P}\in\mathcal{S}} J(\mathbf{P}) \right\}. \tag{4}$$

The goal is to identify the benefit vectors that provide the optimal EJT, so those that maximize the $\psi(\mathbf{P})$ are favorable. In this regard, we consider $\mathbf{P}_\tau^*$ to be the chosen benefit vector that fulfills this condition of optimal EJT in relation to τ, as expressed by:

$$\mathbf{P}_\tau^* \in \arg \max_{\mathbf{P}\in\mathcal{P}_\tau} \psi(\mathbf{P}). \tag{5}$$

Eqs. (4) and (5) satisfy the condition in Definition 2. Therefore, the set of benefit vectors that possess the optimal EJT can be obtained in a procedure that iteratively adjusts the value of τ from its maximum ($\tau_{\max}$) to its minimum ($\tau_{\min}$) in decrements of a step size ε. Given the non-concave nature of $J(\mathbf{P})$ (Sediq et al., 2013), using this method for real-time applications is not practical, especially when the set $\mathcal{S}$ contains a large number of possible benefit vectors. Therefore, finding an alternative solution for such applications is of great importance. In this respect, (Sediq et al., 2013) prove the existence of an alternative equivalent function to obtain $\mathbf{P}_\tau^*$ for any convex set $\mathcal{S}$:

$$\mathbf{P}_\tau^* = \arg \min_{\psi(\mathbf{P})=\tau,\ \mathbf{P}\in\mathcal{S}} \|\mathbf{P}\|^2. \tag{6}$$

Since each element of benefit vectors in $\mathcal{S}$ is bounded by its corresponding safety margin value, $\mathcal{S}$ is considered a hyperrectangle, which makes it a convex set (Boyd & Vandenberghe, 2004). Utilizing (6) in Algorithm 1 enables us to compute the optimal EJT, which, in the scope of this paper, means an efficient fair maximum allocated power for each EVA.

Algorithm 1 Obtaining optimal EJT

```
Input: Distribution system model, operational constraints (1f) to (1h), non-EV loads, time step number $T$, and step $\varepsilon > 0$
Output: $\mathbf{P}^*_\tau$
Initialize grid parameters
for $t = 0$ to $T$ do
    Formulate and solve (1)
    Form the set $S$ using $\mathbf{P}^{SM}$
    Compute $\tau_{\min} = \min_{\mathbf{P} \in S} \psi(\mathbf{P})$, $\tau_{\max} = \max_{\mathbf{P} \in S} \psi(\mathbf{P})$, and $L = \lfloor (\tau_{\max} - \tau_{\min}) / \varepsilon \rfloor$
    for $l = 0$ to $L$ do
        $\tau = \tau_{\max} - \varepsilon l$
        Compute $\mathbf{P}^*_\tau$ using (6)
        if $J(\mathbf{P}^*_\tau) = J(\mathbf{P}^*_{\tau+\varepsilon})$ then
            break
        end if
    end for
end for
return $\mathbf{P}^*_\tau$
```

3.3. MDP formulation of EVAs energy purchasing problem

MDP formulation of EVAs purchasing and distribution of energy within this framework has been inspired by Zhang, Yang, et al. (2023). Accordingly, here are essential notations and definitions for the MDP outlined in this work.

Agents i: in this framework, each EVA is considered as an agent. Here the number of agents is denoted as I.

State Space $\{S_i\}_{i \in I}$: the state space of the environment is denoted as $s_{it} = \{SoC^{av}_{i,t}, N^{EV}_{i,t}, e^g_{i,t}\}$ for each EVA i at a time step t where $SoC^{av}_{i,t}$ is the average SoC of all EVs; $N^{EV}_{i,t}$ is the number of available EVs during charging process; and $e^g_{i,t}$ is the electricity purchasing price for the EVAs which is determined by DSO.

Observation Space $\{O_i\}_{i \in I}$: this work adopts the common assumption that the state of the environment is partially observable for an agent. Each agent i can gain only the observation $o_{it} = \{SoC^{av}_{i,t}, N^{EV}_{i,t}, e^g_{i,t}\}$.

Action Space $\{A_i\}_{i \in I}$: the agent i's discrete action space is specified as $a_{it} = P^g_{i,t}$, with $P^g_{i,t}$ representing the amount of power purchased by agent i from the main grid at time step t.

$$0 \leq P^g_{i,t} \leq P^{\max}_{i,t}. \tag{7}$$

$P^{\max}_{i,t} \in \mathbf{P}^*_\tau$ indicates the maximum power that DSO can dedicate to agent i at time step t.

Transition Dynamic $f(s_{i,t}, a^1_{i,t}, a^2_{i,t}, \ldots, a^n_{i,t}) \rightarrow s_{i,t+1}$ depict the probability of the environment transitioning from state $s_{i,t}$ to $s_{i,t+1}$ when agent i undertake actions $a^1_{i,t}, a^2_{i,t}, \ldots, a^n_{i,t}$.

Reward Function: the reward function in this study takes into account both the profit gained from selling power to the EVs and the cost of purchasing energy from the main grid simultaneously. Hence, in this paper, the reward function for an agent i is defined as follows:

$$r_{it} = (\sum_{z=1}^{Z} P^s_{z,t} \cdot e^s_{i,t} - P^g_{t,i} \cdot e^g_{i,t}) \times \Delta T, \tag{8}$$

where Z is the index of EVs set, $P^s_{z,t}$ denotes the volume of power that an EV z buys from EVA and receives from RSD controller in time step t, $P^g_{t,i}$ is the volume of power that the EVA i purchased from the main grid. Additionally, $e^s_{i,t}$ represents the unit price of power sold by an EVA to EVs, $e^g_{i,t}$ is the unit price of power purchased by an EVA from the main grid, and ΔT represents the duration of each time step.

3.4. Enhanced DDQN

The DDQN represents a reinforcement learning algorithm designed to tackle challenges in sequential decision-making. Unlike the standard DQN, DDQN incorporates two Q-networks to mitigate Q-value overestimation issues (Aljohani, Ebrahim, & Mohammed, 2021). Compared to actor–critic methods like deep deterministic policy gradient (DDPG), which require training both an actor network for policy and a critic network for value estimation, DDQN is computationally less intensive and easier to tune because it only needs to train two similar Q-networks, simplifying the implementation. Furthermore, a prior study by Dorokhova, Martinson, Ballif, and Wyrsch (2021) on RL for EV management highlighted that the DDPG algorithm is highly sensitive to hyperparameter selection, where even a single incorrect parameter can significantly disrupt the learning process. DDQN works well in environments where actions are discrete and the dimensionality of the state space is not excessively high (Nguyen, Nguyen, & Nahavandi, 2020).

The two networks in DQQN serve distinct purposes: one for action selection and the other for Q-value estimation. The choice of actions is determined by the online network based on greedy algorithms, while the estimation of Q-values for the selected actions is carried out by the target network. The target Q value for each agent can be calculated as follows:

$$\begin{aligned} Y_{i,t} = r_{it} + \gamma_i Q_{i,t} (\ & s_{i,t+1}, \\ & \operatorname{argmax}_a Q_{i,t} (s_{i,t+1}, a_{i,t+1}; \theta_{i,t}) ; \theta^-_{i,t}) . \end{aligned} \tag{9}$$

where, for agent i at time step t, the online network's parameters are denoted as $\theta_{i,t}$ while the target network's parameters are denoted as $\theta^-_{i,t}$ and γ_i is the discount factor. Correcting for potential overestimation, the $Q_{i,t}$ value of the primary network is adjusted using $\theta^-_{i,t}$. Following is an expression for the mean square error loss function between the online and target network's Q-value.

$$L(\theta_{i,t}) = E\left[\left(Y_{i,t} - Q_{i,t}(s_{i,t}, a_{i,t}; \theta_{i,t}) \right)^2 \right]. \tag{10}$$

Last, the parameters of the online network are gradually transferred to the parameters of the target network through a slow averaging process with a rate of $\zeta \in (0, 1]$.

$$\theta^-_{i,t} = \zeta \theta_{i,t} + (1 - \zeta) \theta^-_{i,t}. \tag{11}$$

Prioritized experience replay is a methodology that emphasizes the replay of particular experiences within the training process by utilizing a replay buffer (Wang, Tang, Huang, & Wang, 2024). The prioritization is determined by the magnitude of the temporal difference (TD) error δ, which measures the difference between the anticipated Q-value and the actual reward obtained. The representation of the TD error at each time step for each agent is as follows:

$$\begin{aligned} \delta_{i,t} = r_{it} + \gamma Q^{\text{target}}_{i,t} & (s_{i,t+1}, \\ & \operatorname{argmax}_a (Q(s_{i,t}, a_{i,t}))) - Q(s_{i,t}, a_{i,t}) . \end{aligned} \tag{12}$$

The agent's learning efficiency improves when instances characterized by notably large TD errors are replayed more frequently. Experiences become more valuable for learning when a considerable TD error exists indicating a significant difference between the agent's prediction and the actual outcome (Sharma & Thangaraj, 2024). During training, experiences are chosen from the replay buffer based on probabilities linked to their priority values. Events with higher priority values have an elevated likelihood of being sampled and replayed. This process enhances the agent's ability to learn more effectively from its most informative experiences. The probability of sampling the experience tuple d is defined as follows:

$$Z_i(d) = \frac{Pr_{i,d}}{\sum_l Pr_{i,L}}, \tag{13}$$

where $Pr_{i,d}$ represents the priority of experience d for agent i and can be calculated as follows:

$$Pr_{i,d} = |\delta_{i,d}| + \epsilon, \tag{14}$$

where ϵ represents a small positive constant that guarantees that every experience has a chance, however small, of being selected for replay.

Adaptive learning rate refers to a technique in optimization algorithms where the learning rate is adjusted during the training process. The difficulty in selecting an appropriate fixed learning rate can lead to slow optimization, getting stuck in local minima with a too-small rate, or experiencing oscillations and convergence issues with a too-large rate (Yan et al., 2020). The adaptive learning rate enables the algorithm to take larger steps when the optimization is progressing well and smaller steps when it is not. This adaptability can lead to faster convergence (Iiduka, 2022). The following equation indicates the updating process of the learning rate:

$$\alpha_{i,t} = \alpha_i^0 \cdot \frac{\sqrt{1-\beta_{i,t}^2}}{1-\beta_{i,t}^1}, \tag{15}$$

where $(\alpha_{i,t})$ is the adaptive learning rate at time step t for agent i in terms of the initial learning rate (α_i^0), and the hyperparameters $\beta_{i,t}^1$ and $\beta_{i,t}^2$ used in the Adam optimizer.

3.5. Real-time smart dispatch

The objective of the RSD controller is to efficiently distribute the maximum purchased energy among EVs coordinated by EVAs. It begins by sorting EVs based on their charging urgency, and assigning charging priority accordingly. The aim is to prioritize the charging of EVs in a manner that maximizes the energy available for sale to all EVs in the system.

$$\alpha_{i,z,t} = \beta_{\mathrm{SOC}_{i,z,t}} \cdot \frac{d_{z,i}}{(t_{z,i}^{de} - t) \cdot p_{z,i}^{max}}, \tag{16}$$

where at time step t, at EVA i, $\alpha_{i,z,t}$ represent the charging urgency factor of EV z, $d_{i,z,t}$ represents the volume of the electricity demand of EV z, $t_{z,i}^{de} - t$ represents the time interval between the departure time of EV z and the current time, and $p_{z,i}^{max}$ denotes the maximum charging power dedicated to Each EV specifically according to the model of EV z, and finally, $\beta_{\mathrm{SOC}_{i,z,t}}$ represents the SOC-dependent factor which takes into account the lifespan of EV batteries.

$$\beta_{\mathrm{SOC}_{i,z,t}} = e^{-k \cdot \mathrm{SOC}_{i,z,t}}, \tag{17}$$

where $k \in [0,1]$. The exponential decay function indicates that as the SOC increases, the factor $\beta_{\mathrm{SOC}_{i,z,t}}$ decreases exponentially. This implies that there is a diminishing urgency to charge the EV as its SOC increases, resulting in an extended battery lifespan.

In the RSD scheme, a higher urgency in EV charging demand, especially as it approaches departure time or requires more electricity, increases the likelihood of receiving priority in electricity distribution. On the other hand, when EVs reach to an acceptable SoC, the charging rate decreases to prevent any potential harm to the lifespan of the EV batteries. Finally, the SoC of EVs will be updated based on the following formula:

$$\mathrm{SoC}_{i,z,t+1} = \mathrm{SoC}_{i,z,t} + \frac{\Delta t}{C_{i,z}} \cdot P_{i,z,t}^{s}, \tag{18}$$

where at EVA i, $\mathrm{SoC}_{i,z,t}$ is the state of charge of EV z at time t, Δt is the time step, $C_{i,z}$ is the capacity of the battery of EV z, and $P_{i,z,t}^{s}$ is the power consumption of EV z at time t.

Algorithm 2 indicates the enhanced DDQN and RSD scheme. At each episode, during the learning process, agents observe the environment, and actions are executed based on the DDQN network. Following the interaction between agents and RSD controllers, the energy will be dispatched among EVs, the reward will be calculated, and transition tuples will be stored in the replay memory, detailed in lines 8–11. Finally, updates are applied to the enhanced DDQN network and learning rate as depicted in lines 12–15.

Algorithm 2 Enhanced DDQN and RSD scheme Algorithm

```
Input: Episode number $E$, time step number $T$, EVA number $I$
Output: Energy purchasing and distribution strategy for EVAs
 1: Initialize prioritized experience replay $R$
 2: for $e = 1$ to $E$ do
 3:     for $t = 1$ to $T$ do
 4:         for $i = 1$ to $I$ do
 5:             Obtain the observation of agent $o_{it}$
 6:             Obtain $P_{t,i}^{\max}$ from Algorithm 1
 7:             Agent executes action based on $\epsilon$-greedy (7)
 8:             RSD controller dispatch the energy among EVs using
    (16), (17) and report total profit to the agent
 9:             Update the SoC of EVs using (18)
10:             Calculate the reward using (8)
11:             Store transition $(s_{i,t}, A_{i,t}, r_{i,t}, s_{i,t+1})$ in $R$
12:             Sample a mini batch from the replay buffer by taking (13)
    into account
13:             Compute the expected Q-values using (9)
14:             Compute the loss $(L(\theta_{i,t}))$ using (10)
15:             Update target network parameters using (11)
16:         end for
17:         Update the learning rate using (15)
18:     end for
19: end for
20: return The energy purchasing strategy for EVAs
```

Table 1
EVA Information. μ and σ values indicating the mean and standard deviation of arrival/departure times in 15-minute time intervals, respectively.

EVA	EV number	Arrival μ_{arrival}, σ_{arrival}	Departure $\mu_{\text{departure}}$, $\sigma_{\text{departure}}$
1	100	32, 20	64, 16
2	140	30, 16	55, 12
3	150	50, 16	75, 8
4	60	24, 20	50, 8
5	110	36, 12	64, 16
6	80	55, 16	75, 4

4. Case study

4.1. Power system model

In this study, the 118-bus distribution test feeder by Zhang, Fu, and Zhang (2007) is used for the case studies. The system includes six EVAs, strategically positioned at the end of branches. As illustrated in Fig. 2, positioning the EVAs at these specific locations is a deliberate choice, influenced by the fact that buses located at the end of branches encounter more significant challenges related to voltage drops and power losses within the distribution grid. As the framework is designed for online applications, the study considers each time step a 15-minute interval within 24 h. Table 1 presents essential information on the daily inflow of EVs to each EVA, including their arrival and departure times determined by the Gaussian distribution factor. Furthermore, real distribution grids typically accommodate different types of consumers (i.e., residential, commercial, and industrial). To accurately model the grid, demand curve data for different types of consumers is obtained from OpenEI (2014) and Braeuer (2020) representing non-EV loads. These load profiles are then randomly assigned to different buses in the network.

Moreover, to calculate the reward function on (8), a real time pricing scheme has been taken into account as $e_{i,t}^{g}$ which implies the rate of the electricity purchasing price of EVAs from DSO while $e_{i,t}^{s}$ implies a fix rate of 0.5 $/kWh selling energy to EVs. Finally, EVs type and their market share have been extracted from Zhang, Yang, et al. (2023).

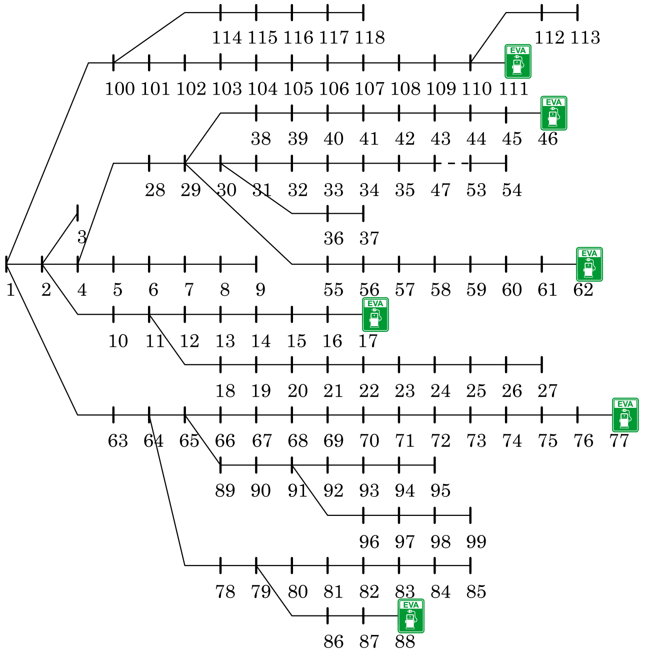


Fig. 2. Distribution network and the location of EVAs.

4.2. Simulation environments and MADRL parameters

By linking MATLAB and Python, this work takes advantage of both environments. Problem (1) is formulated in MATLAB and solved by *fmincon* using MATPOWER (Zimmerman, Murillo-Sánchez, & Thomas, 2011) package. In addition, the MADRL framework is implemented in Python and MATLAB API is used to obtain safe margin and optimal EJT. Using MATLAB API in Python environment renders a real-time simulated framework. The simulation results are obtained on a PC with an Intel Core i7-12700H CPU 2.30 GHz and 40 GB of RAM.

In the context of DDQN parameters, the discount rate γ is set to 0.95, the exploration–exploitation trade-off parameter ϵ in ϵ-greedy is set to 0.95, the mini-batch size is set to 32, the experience replay buffer size is set to 1000, the learning rate is set to 0.001, and the target network update frequency is set to 1000.

5. Experimental results

This section presents the results of three distinct scenarios (SC): 1) the proposed safe MADRL framework (SC1), 2) Uncontrolled charging (SC2) and 3) First come first serve (FCFS) (SC3). A comparative analysis between these three scenarios is then carried out.

5.1. The proposed safe MADRL framework charging (SC1)

In this framework, as explained in Section 2.1, there is an online interaction between the DSO and each EVA at each time step. The maximum allowable power set points are calculated and signal set points are sent to EVAs separately. Accordingly, each EVA (agent) based on its observation determines its energy purchased. In Fig. 3, the daily charging profiles of all EVAs in SC1 are displayed. All EVAs prefer to purchase power during periods of relatively lower electricity prices from the DSO.

Fig. 4 illustrates the charging strategy for EVs associated with EVA 4. The RSD controller allocates energy to each EV based on the urgency of their charging needs at each time step. Initially, upon their arrival, EVs tend to charge at a higher rate with respect to their maximum charging rate, to quickly reach a satisfactory SoC. Once a suitable SoC is achieved, the charging rate decreases to extend the battery's lifespan, gradually leading to the desired SoC.

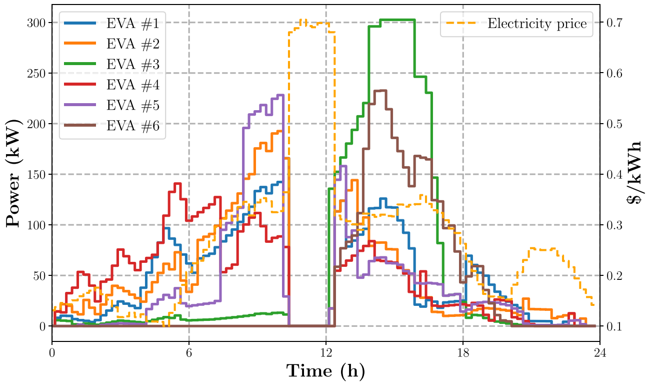


Fig. 3. EVAs daily consumption — SC1.

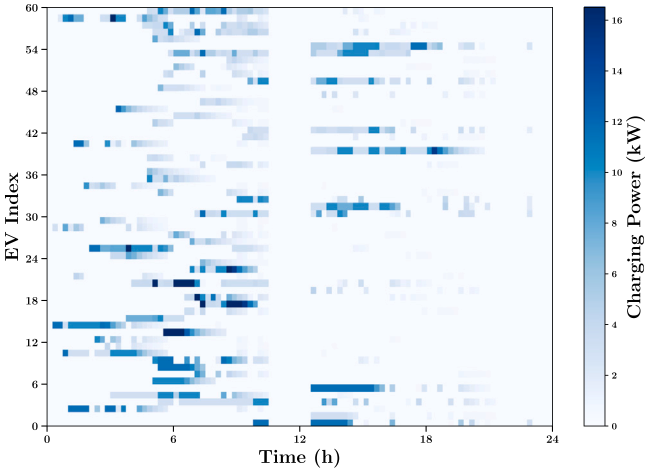


Fig. 4. Charging plan of EVs of EVA #4 - SC1.

Fig. 5 displays the initial and final SoC of EVs of each EVA in SC1. The initial SoC distribution for EVs follows a Gaussian pattern specific to each EVA. The average initial SoC for EVs across all EVAs is 0.45. Fig. 5 also illustrates the final SoC of EVs in SC1, demonstrating the effective performance of RSD in dispatching energy among EVs, given that a significant proportion of the EVs have achieved the desired final SoC.

To indicate the superiority of enhanced DDQN, its performance has been compared with DDQN (Van Hasselt, Guez, & Silver, 2016), DDPG (Lowe et al., 2017) and proximal policy optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017) as illustrated in Fig. 6. The solid and its respected shadow show the mean and standard deviation of the cumulative rewards of all EVAs over 10 runs, respectively. Evidently, enhanced DDQN achieves a near-optimal solution in a significantly reduced number of episodes. A prioritized experience replay helps DDQN focus on more important experiences, enhancing learning efficiency and speeding up convergence to an optimal policy. Additionally, an adaptive learning rate allows for quicker exploration initially and finer adjustments later, improving stability and avoiding suboptimal solutions. These methods combined result in better performance and higher cumulative rewards than standard DDQN, DDPG and PPO. Finally, the computational training times of these four methods are illustrated in Table 2. The enhanced DDQN method took 3134 s, showing a slight increase in time over standard DDQN, which completed in 2569 s. In contrast, DDPG required 4431 s, reflecting the added complexity of its actor–critic architecture. PPO,

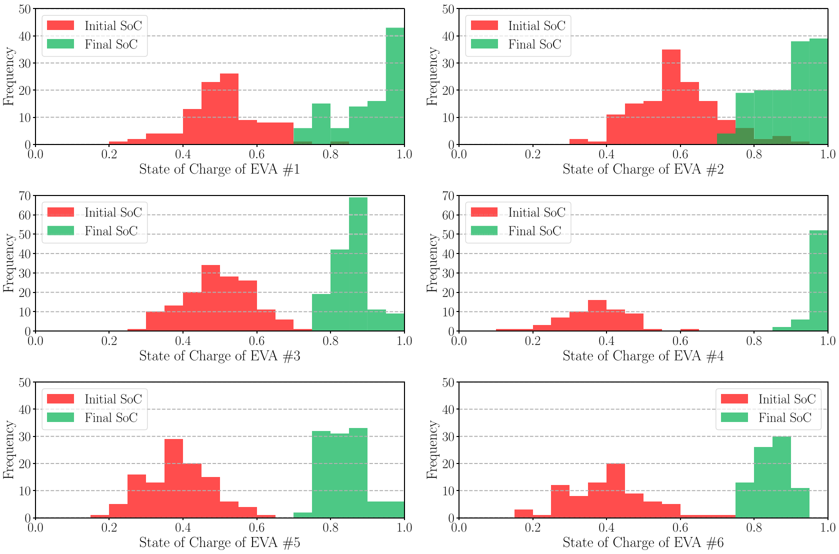


Fig. 5. Initial and final SoC distribution — SC1.

Table 2
Computational training times for different methods in 1000 episodes.

Methods	Enhanced DDQN	DDQN	DDPG	PPO
Computational Training Time (seconds)	3134	2569	4431	5934

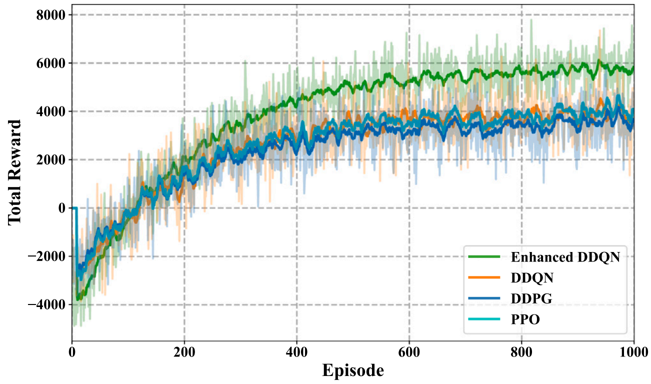


Fig. 6. Comparison of different RL methods over 10 runs — SC1.

being a policy gradient method with stability-enhancing features, exhibited the highest training time at 5934 s, significantly surpassing the other methods.

5.2. Uncontrolled charging (SC2)

In SC2, the assumption is made that all EVs will be charged at their maximum rate upon reaching charging stations. Additionally, it is presumed that there are no restrictions for EVAs regarding purchasing energy from the DSO, allowing them to purchase as much energy as needed according to their demand. As illustrated in Fig. 7, in SC2, the collective charging of a significant number of EVs at their maximum charging rates leads to a higher peak power consumption for EVAs in comparison with SC1 which may cause operational challenges for the DSO. As an example, as shown in Table 3, in SC2, the daily peak power consumption of EVAs 3 and 6 is 498.12 kW and 299.44 kW, respectively. This signifies a 62.22% increase for EVA 3 and a 27.67% increase for EVA 6 when compared to the 307.05 kW and 234.54 kW observed in SC1, respectively.

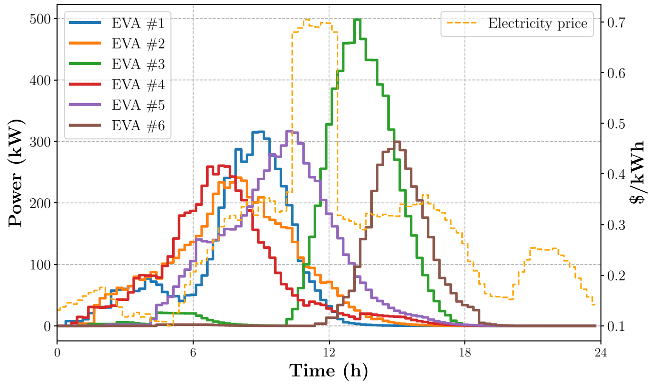


Fig. 7. EVAs daily consumption — SC2.

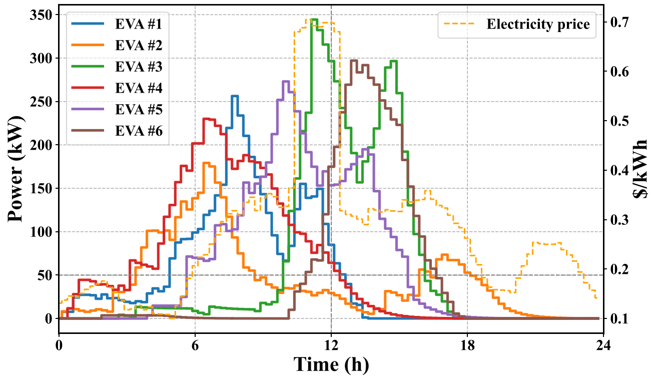
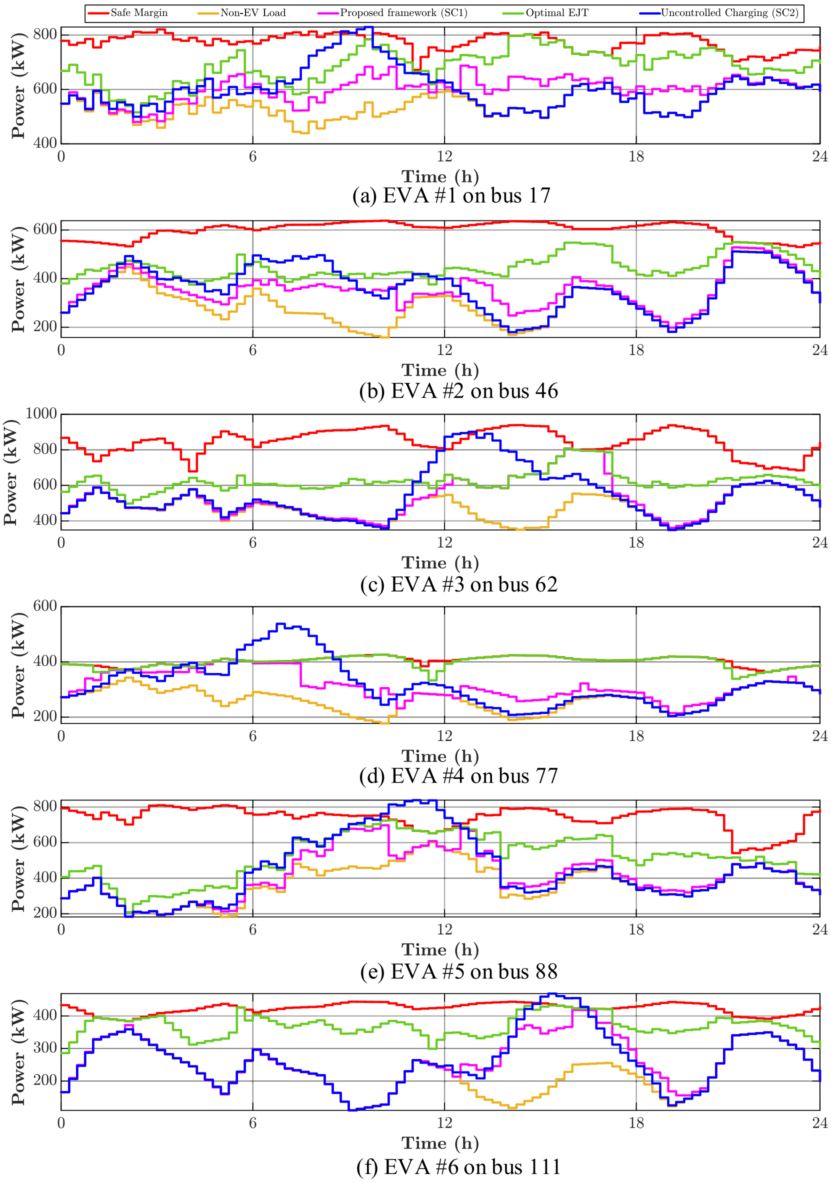


Fig. 8. EVAs daily consumption — SC3.

Due to the fact that in SC2 EVs are charged with their maximum rate of charge, it is undeniable that the final SoC distribution is in better condition than that is in SC1. However, this achievement is obtained at the cost of higher energy purchasing prices for EVAs compared to SC1 as EVAs charge a considerable number of associated EVs during times of on-peak electricity prices as shown in Table 3.



(a) EVA #1 on bus 17

(b) EVA #2 on bus 46

(c) EVA #3 on bus 62

(d) EVA #4 on bus 77

(e) EVA #5 on bus 88

(f) EVA #6 on bus 111

Fig. 9. Power allocation and consumption of EVAs in SC1 and SC2.

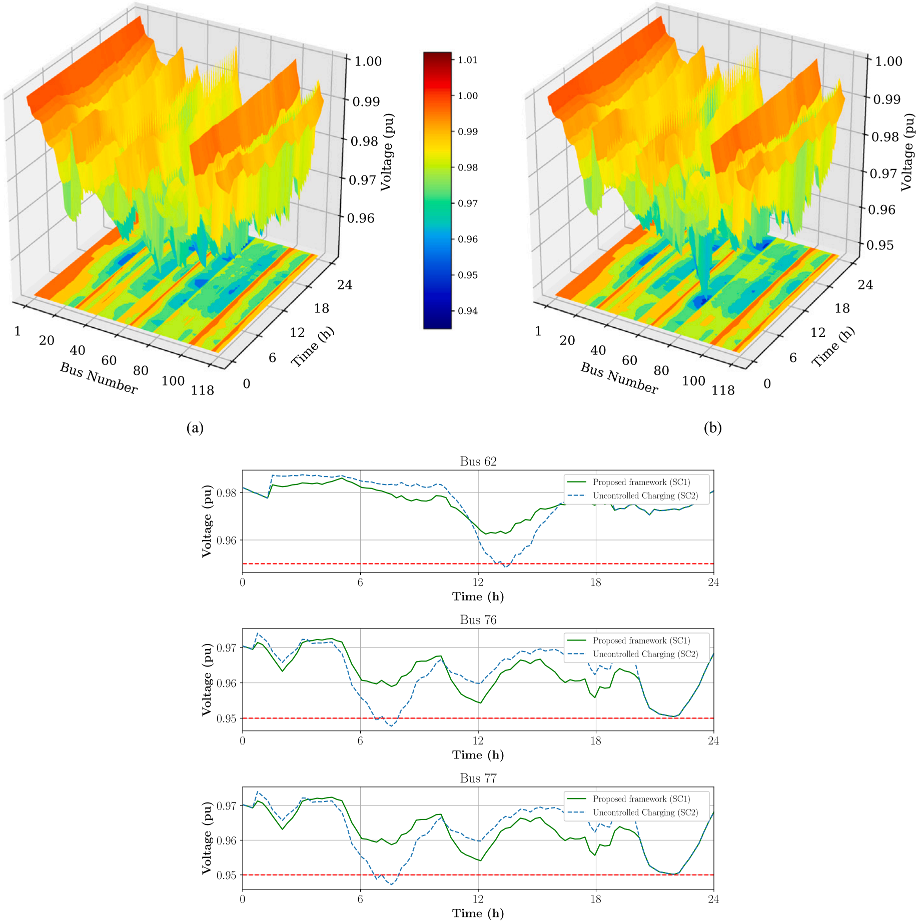
5.3. FCFS (SC3)

In SC3, the assumptions remain identical to those in SC2, with the only difference being the maximum number of EVs allowed for charging per time step, set to 50 per EV aggregator. Additionally, EVs are prioritized based on their arrival times. As illustrated in Fig. 8, the charging profiles of EV aggregators with a larger number of daily EVs are more impacted, as seen with EVAs 3, 5, and 6, where the charging duration is longer compared to SC2. This increase in charging time is due to the prioritization of EVs based on their arrival sequence. Moreover, some EVs may not receive sufficient energy by their departure time since they have waited in the queue for a significant period, which could lead to a decrease in the EV aggregators' profit from charging and cause dissatisfaction among EV owners. Ultimately, for instance, the electricity purchasing costs in SC3 for EVAs 1, 3, and 5 are 1670.11, 2653.54, and 2431.53 $, respectively, representing increases of 9.01%, 19.29%, and 25.66% compared to their corresponding costs in SC1. Additionally, their daily peak power consumptions are 255.82, 346.25, and 271.50 kW, respectively, which are 75.65%, 12.76%, and 19.47% higher than the corresponding daily peak powers in SC1, as depicted in Table 3.

Fig. 9 illustrates power allocation and consumption patterns for EVAs over the 24-hour simulation period. It presents a comparative demonstration of non-EV loads, safe margin values, optimal EJT, SC1, and SC2. The results highlight the effectiveness of the proposed framework in maintaining power consumption within the optimal EJT. This assures the grid's safe operation while considering both efficiency and fairness in power allocation to EVAs, as opposed to uncontrolled charging in SC2. For instance, Fig. 9(a) shows the success of the proposed framework in shifting the EV consumption to avoid peak and distributing the EV demand more evenly across the simulation time. Although SC2 does not violate the safe margin in Fig. 9(b), during some intervals, it exceeds the optimal EJT, suggesting that fairness is not adequately maintained, especially compared to the result observed in SC1. Moreover, despite EVA 3 short window for charging EVs, as per Table 1, Fig. 9(c) indicates its successful performance in using the available time, as indicated by the power consumption closely following the optimal EJT. The correlation between safe margin and optimal EJT for EVA 4, as seen in Fig. 9(d), suggests that the grid constraints are tight at bus 77. Therefore, considering the fairness concept enables EVA 4 to potentially be allowed to draw maximum power (i.e., safe margin) over many intervals. Lastly, Fig. 9(e) and Fig. 9(f) display the effectiveness of

Table 3
Electricity Purchasing Costs and Peak Power for EVA.

EVA	Peak power in SC1 (kW)	Electricity purchasing cost in SC1 ($)	Peak power in SC2 (kW)	Electricity purchasing cost in SC2 ($)	Peak power in SC3 (kW)	Electricity purchasing cost in SC3 ($)
EVA #1	**145.64**	**1532.10**	**315.73**	**2127.44**	**255.82**	**1670.11**
EVA #2	**188.28**	**1701.76**	**241.05**	**1805.48**	**178.15**	**1203.67**
EVA #3	**307.05**	**2224.48**	**498.12**	**2559.07**	**346.25**	**2653.54**
EVA #4	**142.11**	**1172.06**	**260.50**	**1345.69**	**230.19**	**1678.92**
EVA #5	**227.25**	**1935.02**	**316.55**	**3129.68**	**271.50**	**2431.53**
EVA #6	**234.54**	**1935.02**	**299.44**	**1663.02**	**297.45**	**1688.13**



(a)

(b)

(c)

Fig. 10. Voltage magnitude. (a) Voltage under proposed framework (SC1). (b) Voltage under uncontrolled charging (SC2). (c) Buses with voltage violation.

the proposed framework in charging EVs in a shiftable manner without overloading the grid even with a late average arrival time for EVs, as indicated in Table 1. Moving from analyzing EVAs' power consumption patterns, Fig. 10 presents an examination of voltage profiles under two scenarios. In this study, the acceptable voltage magnitude for buses is defined with a lower bound of 0.95 pu and an upper bound of 1.05 pu. As depicted in Fig. 10(a), voltage of buses in SC1 is within the acceptable range. Yet, in SC2, violations of this constraint are observed at buses 62, 76, and 77, as shown in Fig. 10(b) and Fig. 10(c). At bus 62, SC1 successfully maintains the voltage level within the safe operational limits. This, in fact, is obtained by allocating a value below safe margin power (i.e., optimal EJT) to the EVA. Conversely, under SC2, a deviation in voltage is observed, which matches the EVA 3 peak charging period in Fig. 9(c). Furthermore, the results at buses 76 and 77 highlight the benefit of using the MADRL approach to manage EV charging through EVA. Given that no EVA is connected to bus 76, the results suggest that uncontrolled charging may lead to unsafe grid operation. It is also observed that among the buses with EVAs, only buses 62 and 77 experience a voltage drop under SC2. While EVA 2 in SC2 stays within the safe margin, the EVAs associated with buses 17, 88, and 111 distinctly exceed their respective safe margins. This indicates that another critical operational threshold of the grid, line capacity or thermal limit, is at risk.

Fig. 11 illustrates the allocated power to each EVA during the 24-hour period obtained by optimal EJT calculation. As it is demonstrated in Fig. 9(d), allocated power to EVA 4 is mostly equal to the safe margin. This means that at certain times during the day, EVA 4 has a lower allocated power compared to other EVAs.

5.4. Integration of RESs

In this section, we investigate the performance of the proposed framework in the presence of RESs, focusing on PV systems. The location of these systems in the grid affects the allocated power to each EVA and potentially the safe margin. We first demonstrate how PV integration into buses with existing EVAs mainly alters the allocated

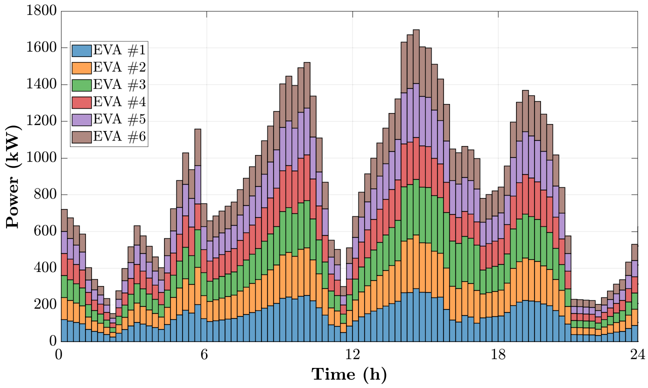


Fig. 11. Allocated power to each EVA by DSO through optimal EJT calculation.

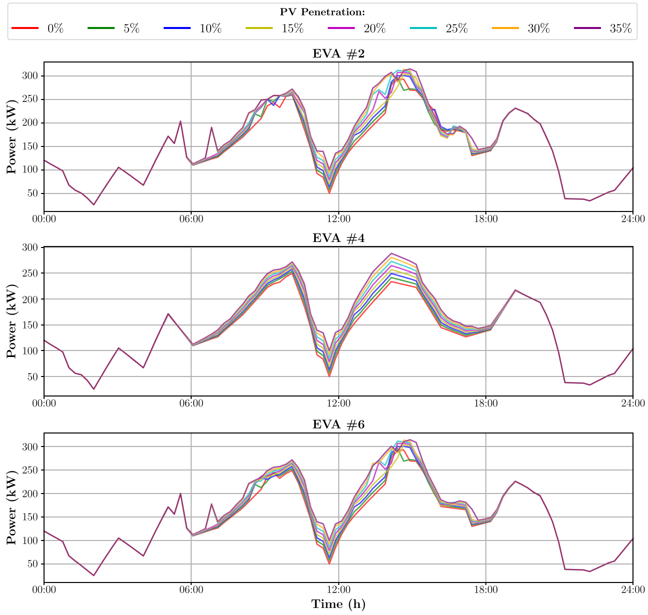


Fig. 12. Allocated power to each EVA by DSO through optimal EJT calculation under different PV penetration.

power to those EVAs. Second, we show that installing PV systems on other buses can potentially enhance the safe margin value.

5.4.1. PV placement on EVA buses

In this subsection, we analyze the effects of placing PV systems on buses with EVAs. In this scenario, it is assumed that PV systems are installed on all six buses hosting EVAs. A sensitivity analysis is then performed based on the percentage of total energy consumed by each EVA throughout the day. Specifically, we evaluate cases where the PV systems can supply up to 35 percent of the total energy required by an EVA over a 24-hour period. The safe margin is tied to the constraints of the grid and may not be significantly changed by the placement of PV systems under typical operating conditions. However, using PV allows the bus to satisfy part of its demand locally. As all EVAs benefit from this opportunity, we should expect an increase in the power allocated to EVAs derived from optimal EJT during the daytime when solar radiation is available, as shown in Fig. 12. This increase makes the grid more reliable and helps EVAs potentially earn more money by relying less on costly grid electricity.

5.4.2. PV placement on other buses

In this scenario, PVs are installed at buses 41, 69, and 105. These locations are chosen because they are positioned near the middle of the branches, with EVAs located at the branch ends. This placement can potentially help mitigate voltage drops and reduce loading on the lines. To investigate the impact of PV penetration levels on the safe margin values, a sensitivity analysis is performed by changing the percentage of the bus's total energy demand that is supplied by PV during the day. As the results demonstrate in Fig. 13, regardless of the level of PV penetration at bus 41, the safe margin values of EVA in the same branch do not change. The ongoing congestion in the branch connecting buses 41 and 46 is causing this issue. This also indicates the need for corrective actions, such as enhancing lines' capacity. On the other hand, PV penetration on buses 69 and 104 effectively increase the safe margin values at the EVAs at the end of their respective branches.

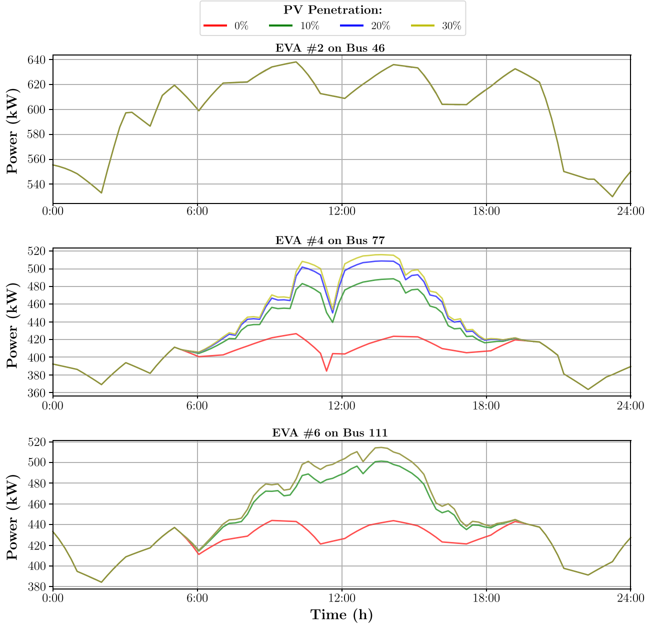


Fig. 13. PSafe margin variations for different PV penetration levels with installations located at the midpoints of branches.

6. Conclusion

This research introduced a MADRL framework for managing energy purchasing and distribution for EVAs while considering distribution network constraints. The framework operates hierarchically, allocating maximum allowable power to each EVA to ensure grid safety, while also achieving equitable energy distribution among EVAs. At the top level, it identifies the optimal EJT point, balancing the maximum total energy the DSO can sell with equitable power allocation among the EVAs.

At the lower level, each EVA functions as an autonomous agent, adopting a DDQN with adaptive learning rates and prioritized experience replay. Through this method, EVAs are able to refine their energy purchasing strategies, increasing their profits. Additionally, at this level, an RSD controller manages the energy distribution among EVs based on their requirements. The proposed framework, including scenarios with PV systems integrated at various buses, is implemented on the 118-bus distribution test feeder, and its performance is compared with uncontrolled and FCFS charging scenarios. The results suggest that the proposed framework significantly enhances grid stability and energy distribution, outperforming the two other mentioned scenarios in terms of both energy purchasing price and peak demand reduction for EVAs.

As a next step in this research, the integration of uncertainty in energy demand and EV behavior should be explored, along with the application of robustness techniques to ensure reliable system performance during grid faults or price fluctuations. Subsequent works can also explore MARL approaches that consider interactions between EV aggregators and other grid elements such as energy markets.

CRediT authorship contribution statement

Arian Shah Kamrani: Software simulation, Methodology, Formal Analysis, Writing – original draft. **Anoosh Dini:** Data Extracting, Formal Analysis, Software simulation, Writing. **Hanane Dagdougui:** Supervision, Writing – review & editing. **Keyhan Sheshyekani:** Supervision, Writing – review & editing.

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.

Data availability

Data will be made available on request.

References

Abdullah, H. M., Gastli, A., & Ben-Brahim, L. (2021). Reinforcement learning based EV charging management systems–A review. *IEEE Access*, *9*, 41506–41531.

Aljohani, T. M., Ebrahim, A., & Mohammed, O. (2021). Real-Time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and Markov chain model. *Electric Power Systems Research*, *192*, Article 106962.

Amini, M. H., McNamara, P., Weng, P., Karabasoglu, O., & Xu, Y. (2018). Hierarchical electric vehicle charging aggregator strategy using dantzig-wolfe decomposition. *IEEE Design & Test*, *35*(6), 25–36.

Arwa, E. O., & Folly, K. A. (2020). Reinforcement learning techniques for optimal power control in grid-connected microgrids: A comprehensive review. *IEEE Access*, *8*, 208992–209007.

Bai, J., Ding, T., Jia, W., Zhu, S., Bai, L., & Li, F. (2024). Online rectangle packing algorithm for swapped battery charging dispatch model considering continuous charging power. *IEEE Transactions on Automation Science and Engineering*, *21*(1), 320–331.

Bao, Z., Hu, Z., & Mujeeb, A. (2024). A novel electric vehicle aggregator bidding method in electricity markets considering the coupling of cross-day charging flexibility. *IEEE Transactions on Transportation Electrification*, 1.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization* (pp. 21–66). Cambridge University Press.

Braeuer, F. (2020). *Load profile data of 50 industrial plants in Germany for one year*. Zenodo.

Cao, Y., Wang, H., Li, D., & Zhang, G. (2022). Smart online charging algorithm for electric vehicles via customized actor–critic learning. *IEEE Internet of Things Journal*, *9*(1), 684–694.

Ding, T., Zeng, Z., Bai, J., Qin, B., Yang, Y., & Shahidehpour, M. (2020). Optimal electric vehicle charging strategy with Markov decision process and reinforcement learning technique. *IEEE Transactions on Industry Applications*, *56*(5), 5811–5823.

Dorokhova, M., Martinson, Y., Ballif, C., & Wyrsch, N. (2021). Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation. *Applied Energy*, *301*, Article 117504.

Hupez, M., Toubeau, J.-F., De Grève, Z., & Vallée, F. (2021). A new cooperative framework for a fair and cost-optimal allocation of resources within a low voltage electricity community. *IEEE Transactions on Smart Grid*, *12*(3), 2201–2211.

Hussain, A., & Musilek, P. (2022). Fairness and utilitarianism in allocating energy to EVs during power contingencies using modified division rules. *IEEE Transactions on Sustainable Energy*, *13*(3), 1444–1456.

Iiduka, H. (2022). Appropriate learning rates of adaptive learning rate optimization algorithms for training deep neural networks. *IEEE Transactions on Cybernetics*, *52*(12), 13250–13261.

Jain, R. K., Chiu, D.-M. W., Hawe, W. R., et al. (1984). *A quantitative measure of fairness and discrimination*: *vol. 21*, Hudson, MA: Eastern Research Laboratory, Digital Equipment Corporation.

Jin, R., Zhou, Y., Lu, C., & Song, J. (2022). Deep reinforcement learning-based strategy for charging station participating in demand response. *Applied Energy*, *328*, Article 120140.

Kiani, S., Sheshyekani, K., & Dagdougui, H. (2023). An extended state space model for aggregation of large-scale EVs considering fast charging. *IEEE Transactions on Transportation Electrification*, *9*(1), 1238–1251.

Kiani, S., Sheshyekani, K., & Dagdougui, H. (2024). ADMM-based hierarchical single-loop framework for EV charging scheduling considering power flow constraints. *IEEE Transactions on Transportation Electrification*, *10*(1), 1089–1100.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, *30*.

Madahi, S. S. K., Kamrani, A. S., & Nafisi, H. (2022). Overarching sustainable energy management of PV integrated EV parking lots in reconfigurable microgrids using generative adversarial networks. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 19258–19271.

Nguyen, T. T., Nguyen, N. D., & Nahavandi, S. (2020). Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, *50*(9), 3826–3839.

OpenEI (2014). Commercial and residential hourly load profiles for all TMY3 locations in the United States. https://openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states.

Paudel, D., & Das, T. K. (2023). A deep reinforcement learning approach for power management of battery-assisted fast-charging EV hubs participating in day-ahead and real-time electricity markets. *Energy*, *283*, Article 129097.

Poudel, S., Mukherjee, M., Sadnan, R., & Reiman, A. P. (2023). Fairness-aware distributed energy coordination for voltage regulation in power distribution systems. *IEEE Transactions on Sustainable Energy*, *14*(3), 1866–1880.

Qi, C., Liu, C.-C., Lu, X., Yu, L., & Degner, M. W. (2023). Transactive energy for EV owners and aggregators: Mechanism and algorithms. *IEEE Transactions on Sustainable Energy*, *14*(3), 1849–1865.

Saner, C. B., Trivedi, A., & Srinivasan, D. (2022). A cooperative hierarchical multi-agent system for EV charging scheduling in presence of multiple charging stations. *IEEE Transactions on Smart Grid*, *13*(3), 2218–2233.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Sediq, A. B., Gohary, R. H., Schoenen, R., & Yanikomeroglu, H. (2013). Optimal tradeoff between sum-rate efficiency and jain's fairness index in resource allocation. *IEEE Transactions on Wireless Communication*, *12*(7), 3496–3509.

Sharma, A., & Thangaraj, V. (2024). Intelligent service placement algorithm based on DDQN and prioritized experience replay in IoT-Fog computing environment. *Internet of Things*, *25*, Article 101112.

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 30, no. 1*.

Wang, Q., Chen, Z., Zhou, Y., Liu, Z., & Peng, Z. (2024). Real-time monitoring and optimization of machine learning intelligent control system in power data modeling technology. *Machine Learning with Applications*, *18*, Article 100584.

Wang, B., Tang, Y., Huang, Y., & Wang, T. (2024). Power system emergency control strategy based on severely disturbed units identification and STGCN-DDQN. *Electric Power Systems Research*, *226*, Article 109903.

Wu, D., Radhakrishnan, N., & Huang, S. (2019). A hierarchical charging control of plug-in electric vehicles with simple flexibility model. *Applied Energy*, *253*, Article 113490.

Yan, X., Xu, Y., Xing, X., Cui, B., Guo, Z., & Guo, T. (2020). Trustworthy network anomaly detection based on an adaptive learning rate and momentum in IIoT. *IEEE Transactions on Industrial Informatics*, *16*(9), 6182–6192.

Yuan, C., Forhad, M. A. A., Bansal, R., Sidorova, A., & Albert, M. V. (2024). Multi-agent dual level reinforcement learning of strategy and tactics in competitive games. *Results in Control and Optimization*, *16*, Article 100471.

Zhang, D., Fu, Z., & Zhang, L. (2007). An improved TS algorithm for loss-minimum reconfiguration in large-scale distribution systems. *Electric Power Systems Research*, *77*(5), 685–694.

Zhang, C., Liu, Y., Wu, F., Tang, B., & Fan, W. (2021). Effective charging planning based on deep reinforcement learning for electric vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *22*(1), 542–554.

Zhang, Y., Rao, X., Liu, C., Zhang, X., & Zhou, Y. (2023). A cooperative EV charging scheduling strategy based on double deep Q-network and prioritized experience replay. *Engineering Applications of Artificial Intelligence*, *118*, Article 105642.

Zhang, Y., Yang, Q., An, D., Li, D., & Wu, Z. (2023). Multistep multiagent reinforcement learning for optimal energy schedule strategy of charging stations in smart grid. *IEEE Transactions on Cybernetics*, *53*(7), 4292–4305.

Zheng, Y., Song, Y., Hill, D. J., & Meng, K. (2019). Online distributed MPC-based optimal scheduling for EV charging stations in distribution systems. *IEEE Transactions on Industrial Informatics*, *15*(2), 638–649.

Zimmerman, R. D., Murillo-Sánchez, C. E., & Thomas, R. J. (2011). MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, *26*(1), 12–19.