Deep RL-Based Operation of Clustered EV Charging Stations Integrated with PV and Battery

Shima Tayefi
Electrical and Computer Engineering
Laval University
Quebec, Canada
shima.tayefi.1@ulaval.ca

Van-Hai Bui
Electrical and Computer Engineering
University of Michigan-Dearborn
Michigan, USA
vhbui@umich.edu

Akhtar Hussain
Electrical and Computer Engineering
Laval University
Quebec, Canada
akhtar.hussain@gel.ulaval.ca

***Abstract*—Rapid electric vehicle (EV) adoption is stressing local distribution systems due to clustered charging and limited transformer capacity. Although vehicle-to-grid (V2G) coordination and distributed energy resource (DER) integration offer additional flexibility, their operation remains challenging under uncertain EV behaviour and variable renewable output. Therefore, this study analyzes two state-of-the-art deep reinforcement learning algorithms (soft actor-critic (SAC) and twin delayed deep deterministic policy gradient (TD3)) to minimize the operation cost of a residential charging station equipped with PV, a community battery, and a charging management system. The learned strategies are benchmarked against a mixed-integer linear programming (MILP) formulation used as the optimal reference. Simulation results demonstrate the effectiveness of both methods, with SAC achieving performance closest to the optimal solution across diverse EV load scenarios.**

Keywords: Battery energy storage, deep reinforcement learning, electric vehicles, energy management, SAC, TD3

I. Introduction

The size of the global electric vehicle (EV) fleet has increased steadily over the last decade, driven by environmental concerns, government incentives, and rapid advancements in battery technologies. However, this widespread adoption introduces several technical challenges for modern power systems, particularly at the distribution level where infrastructure is often not designed for high and simultaneous charging demand [1]. Large numbers of EVs connected within the same locality can create significant stress on distribution transformers and feeder lines, leading to voltage deviations, thermal overloads, and accelerated asset aging. These issues are especially pronounced in clustered charging environments (residential apartment complexes with shared parking, university campuses, commercial buildings, and schools), where multiple EVs may charge concurrently during similar time windows [2].

To address these issues, numerous studies have been conducted in recent years [3], [4], and the existing approaches can be broadly classified into two categories. The first category focuses on local demand management of EV charging, where decentralized or semi-decentralized control mechanisms regulate charging based on locally available information or market-like frameworks. These methods often incorporate local EV coordination, peer-to-peer (P2P) energy exchange, or localized pricing schemes designed to balance charging demand among EVs without requiring extensive grid-level communication. For example, A network distributed algorithm has been proposed to enable peer-to-peer coordination among charging stations for preventing node congestion in [5]. Similarly, [6] introduced a semi-distributed control mechanism that supports local decisions, satisfies charging demand, and tracks renewable generation. The pros and cons of local and centralized strategies in terms of capacity utilization and total energy delivered to charging vehicles has been analyzed in [7] and showed that local control can maximize charging rate while maintaining grid conditions.

Despite their advantages, local EV coordination and P2P schemes require active EV participation and data sharing, which raises privacy concerns and depends heavily on user willingness. Moreover, these approaches typically overlook key distribution-network constraints (transformer limits, feeder loading, and voltage drops). To overcome these drawbacks, a second category of research focuses on integrating local distributed energy resources (DERs), such as PV, energy storage system (ESS), and building-level energy management systems, to support clustered EV charging while explicitly accounting for grid constraints. For instance, a decentralized model predictive control-based control method for smart charging of electric vehicles and their coordination with photovoltaic systems has been proposed in [8]. Similarly, optimal scheduling of interactive charging for electric vehicles in solar-powered parking lots, considering uncertainties and vehicle-to-grid capability has been analyzed in [9]. Finally, integration of renewable energy and electric vehicles in power systems was reviewed in [10].

However, DER-based coordination remains challenging, largely due to the inherent uncertainty in predicting EV charging demand and PV generation. Highly variable user behavior, irregular arrival times, inconsistent charging needs, and diverse energy preferences, makes EV load estimation difficult [11]. Meanwhile, PV output is equally sensitive to weather and rapidly changing local conditions. These uncertainties can undermine scheduling performance and limit the effectiveness of conventional control approaches. Consequently, there is a growing need for robust, data-driven, and adaptive strategies, such as deep reinforcement learning (DRL), which can dynamically learn optimal charging policies and respond to stochastic

979-8-3315-8138-1

EV and renewable patterns [12], [13].

To address the limitations identified in earlier studies, this work investigates two advanced DRL algorithms, soft actor-critic (SAC) and twin delayed deep deterministic policy gradient (TD3) for managing the energy flows in a residential charging station equipped with PV generation and an ESS. These algorithms are selected due to their proven stability and strong performance in continuous control tasks. The considered system includes a rooftop PV, an ESS, and a shared parking facility with several EVs. All resources are supervised by a charging management system (CMS) that schedules charging, discharging, and grid interactions to reduce operating costs while respecting equipment and operational constraints. The objective is to learn effective control policies using SAC and TD3, and their performance is compared with a benchmark mixed-integer-linear-programming (MILP) formulation.

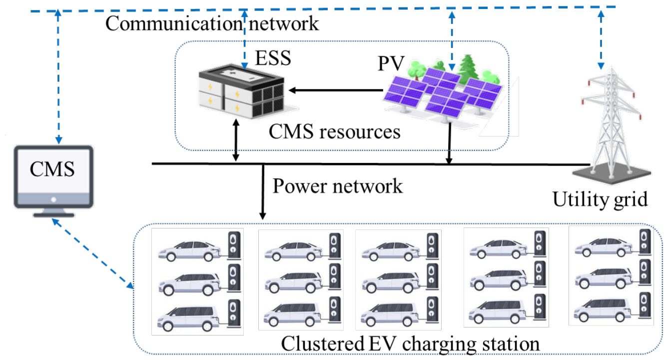

Fig. 1. Overview of the system considered in this study.

II. Proposed DRL-Based Operation of CMS

A. System Configuration

The system under study consists of a residential shared parking lot equipped with local photovoltaic (PV) generation and an ESS, as shown in Fig. 1. In this study, a small-scale, single-site microgrid is modeled; therefore, power-flow losses are neglected and voltage and line-current constraints are assumed to be satisfied. A charging management system (CMS) coordinates the charging activities of all connected EVs at the site. The CMS has access to external grid's price signals and schedules the local resources (ESS, PV generation, and grid exchange) while respecting local equipment (transformer and charger) limitations. The proposed DRL-based control strategies (discussed in the subsequent section) are implemented within the CMS, enabling adaptive and optimal management of EV charging.

B. DRL-Based CMS Operation

The operation of the CMS within the proposed system configuration is realized using two state-of-the-art DRL algorithms (SAC and TD3). These algorithms are chosen for their robustness, stability, and strong performance in continuous control problems. By leveraging their actor–critic architectures, the CMS learns optimal charge/discharge decisions over a 24-hour horizon while accounting for variations in PV generation, EV charging demand, and energy prices. Details about the state space, action space, and reward formulation used to train both algorithms are discussed below.

1) State and Action Space

The state vector incorporates the information required for charging station operation, including battery state of charge (SOC), PV generation, EV charging demand, and time-ofuse (TOU) electricity prices. At each time step $t \in \{0, 1, \dots, T-1\}$, the state is defined as:

$$s_t = \left(\mathrm{SOC}_t,\ Pr_t^{\text{buy}},\ Pr_t^{\text{sell}},\ P_t^{\text{pv}},\ P_t^{\text{ev}},\ \tilde{a}_{t-1}\right) \in \mathbb{R}^6 \tag{1}$$

where $\mathrm{SOC}_t \in [0.1, 0.9]$ is the battery SOC, Pr_t^{buy} and Pr_t^{sell} are the electricity purchase and selling prices ($/kWh), P_t^{pv} is PV power (kW), P_t^{ev} is EV demand (kW) and $\tilde{a}_{t-1} \in [-1, 1]$ is the previous normalized action. This normalized value is then linearly mapped to the ESS power as:

$$a_t = \begin{cases} \tilde{a}_t\, P_{\text{ch}}^{\max}, & \text{if } \tilde{a}_t \geq 0, \\ \tilde{a}_t\, P_{\text{dis}}^{\max}, & \text{if } \tilde{a}_t < 0, \end{cases} \tag{2}$$

where $a_t \in [-P_{\text{dis}}^{\max}, P_{\text{ch}}^{\max}]$ represents the BESS power. The actual charging and discharging powers are obtained as:

$$P_t^{\text{ch}} = \max\big(a_t, 0\big), \qquad P_t^{\text{dis}} = \max\big(-a_t, 0\big). \tag{3}$$

2) Constraints

To ensure smooth operation and limit power ramping, the change in battery power between consecutive time steps is limited:

$$|a_t - a_{t-1}| \leq \Delta P_{\text{ramp}} \tag{4}$$

where ΔP_{ramp} is the maximum allowable power change per time step. This constraint is enforced by clipping the physical action after mapping from the normalized action $\tilde{a}_t$:

$$a_t \leftarrow \text{clip}\left(a_t,\ a_{t-1} - \Delta P_{\text{ramp}},\ a_{t-1} + \Delta P_{\text{ramp}}\right) \tag{5}$$

At every simulation step, the active power balance is monitored and the grid import/export variables are adjusted as needed to keep power balance in the system.

3) Reward Function

The reward function is designed to maximize the operational profit of the charging station by optimally coordinating energy exchanges among the ESS, PV, and the utility grid. The instantaneous reward consists of three components. The first component represents the profit of trading energy with the utility grid. The second component penalizes violations of the battery operating range (e.g., SOC or power limits). The third component penalizes any unmet EV charging demand. The total reward is calculated as:

$$\begin{aligned} R_{\text{total}} = \sum_{t=0}^{T-1} \kappa \Big\{ &\left(Pr_t^{\text{sell}} P_t^{\text{sell}} - Pr_t^{\text{buy}} P_t^{\text{buy}}\right) \\ &- \lambda_{\text{bat}}\, \Phi_{\text{batt},t} - \lambda_{\text{EV}}\, \Phi_{\text{EV},t} \Big\} \end{aligned} \tag{6}$$

$$\Phi_{\text{batt},t} = \begin{cases} 1, & \text{if } \mathrm{SOC}_t > 0.9 \text{ or } \mathrm{SOC}_t < 0.1, \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

979-8-3315-8138-1

where $\kappa > 0$ is the reward scaling factor, λ_{batt} and λ_{EV} are penalty coefficients for battery constraint violations and unmet EV load, respectively. Finally, $\Phi_{\text{batt},t}, \Phi_{\text{EV},t} \geq 0$ denote the corresponding penalty functions at time step t.

Each episode spans a 24-hour horizon. At the final time step, a terminal SOC requirement $\text{SOC}_{\text{target}}$ may be enforced. In such cases, the reward at the terminal step is adjusted by a quadratic penalty reflecting the deviation between the actual terminal SOC and the target value, i.e.,

$$r_{T-1} \leftarrow r_{T-1} - \lambda_{\text{terminal}}\big(\text{SOC}_T - \text{SOC}_{\text{target}}\big)^2 \tag{8}$$

where $\lambda_{\text{terminal}} \geq 0$ is the terminal SOC penalty coefficient. This reward formulation encourages the agent to minimize energy costs throughout the day while maintaining a desirable SOC at the end of each episode.

C. SAC-Based Realization

The SAC algorithm learns a stochastic policy by maximizing an entropy-regularized objective that balances cost minimization and exploration, [14]:

$$\mathcal{J}(\phi) = \mathbb{E}_{\tau\sim\pi_\phi}\left[\sum_{t=0}^{T-1} \gamma^t \left(r_t + \alpha\mathcal{H}(\pi_\phi(\cdot|s_t))\right)\right] \tag{9}$$

where γ is the discount factor, α is the temperature parameter, and $\mathcal{H}(\pi) = -\mathbb{E}_{a\sim\pi}[\log \pi(a)]$ is policy entropy [14].

Policy network: The stochastic actor is modeled as a Gaussian policy, as given in $\mu_\phi(s)$ and standard deviation $\sigma_\phi(s)$.

$$\pi_\phi(a|s) = \mathcal{N}(a \mid \mu_\phi(s)\sigma_\phi^2(s)) \tag{10}$$

$$a = \tanh\left(\mu_\phi(s) + \sigma_\phi(s) \odot \epsilon\right), \quad \epsilon \sim \mathcal{N}(0, I) \tag{11}$$

where $\tanh(\cdot)$ squashes actions to $[-1, 1]$.

The actor network contains two-layer MLP with 256 ReLU units per layer, followed by linear heads generating $\mu_\phi(s)$ and $\log \sigma_\phi(s)$. Orthogonal initialization is used for all layers.

Critic networks: Two Q-functions $Q_{\theta_1}(s, a)$ and $Q_{\theta_2}(s, a)$ mitigate overestimation bias, as expressed by :

$$Q_{\theta_i}(s, a) : \mathbb{R}^6 \times [-1, 1] \to \mathbb{R}, \quad i \in \{1, 2\}.$$

Each critic receives the concatenated state–action vector (s, a) as input and is implemented as a two-layer MLP with 256 hidden units per layer.

D. TD3-Based Realization

The TD3 algorithm is an improved deterministic actor–critic algorithm designed to reduce overestimation and enhance stability. TD3 incorporates three mechanisms, including twin critics, target policy smoothing, and delayed actor updates to improve convergence in continuous-control settings such as ESS dispatch.

The objective is to learn a deterministic policy $\pi_\phi(s)$ that maximizes the expected discounted return, as given as:

$$\mathcal{J}_{\text{return}}(\phi) = \mathbb{E}_{\mathcal{T}\sim\pi_\phi}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right] \tag{12}$$

where $\gamma \in (0, 1)$ is the discount factor, $\mathcal{T}$ denotes a trajectory, and we use ϕ to denote the deterministic policy parameters.

Policy and critic networks: The deterministic actor is defined as:

$$a_t = \pi_\phi(s_t) = \tanh\big(\mu_\phi(s_t)\big) \tag{13}$$

where μ_ϕ is a neural network mapping states directly to unnormalized actions. The $\tanh(\cdot)$ squashes the output to the normalized range $[-1, 1]$, which is then linearly scaled to the physical power range $[-P_{\max}^{\text{dis}}, P_{\max}^{\text{ch}}]$.

TD3 employs two critic networks , $Q_{\theta_1}(s, a)$ and $Q_{\theta_2}(s, a)$, each receiving the concatenated state–action vector (s, a) and sharing the same two-layer 256-unit MLP structure. The twin-critic design mitigates Q-value overestimation.

Target policy smoothing: TD3 introduces clipped noise to the target action to improve robustness as :

$$a' = \pi_{\phi'}(s) + \varepsilon, \qquad \varepsilon \sim \text{clip}\big(\mathcal{N}(0, \sigma^2), -c, c\big) \tag{14}$$

where σ and c denote the noise standard deviation and the clipping threshold, respectively. This smoothing prevents the policy from exploiting sharp or unrealistically optimistic regions of the Q-function.

E. Training Procedure

Both SAC and TD3 follow an off-policy actor–critic training paradigm with a replay buffer to stabilize learning. At each training iteration, a minibatch of transitions $(s_t, a_t, r_t, s_{t+1}, d_t)$ is sampled from the buffer to update the critic and actor networks.

Critic update: For both algorithms, the critic parameters are updated by minimizing the temporal-difference (TD) loss between the current Q-value estimates and their respective targets, as defined in (15) for SAC and (16) for TD3.

$$\begin{aligned} y^{\text{SAC}} = r + \gamma(1-d)\Big(&\min_{i\in\{1,2\}} Q_{\theta_i}(s', a') \\ &- \alpha \log \pi_\phi(a' \mid s')\Big) \end{aligned} \tag{15}$$

$$y^{\text{TD3}} = r + \gamma(1-d) \min_{i\in\{1,2\}} Q_{\theta_i}(s', a'). \tag{16}$$

where $a' \sim \pi_\phi(\cdot \mid s')$ for SAC, and $a' = \pi_{\phi'}(s') + \varepsilon$, $\varepsilon \sim \text{clip}\big(\mathcal{N}(0, \sigma^2), -c, c\big)$ for TD3, which implements target policy smoothing. Each critic is trained by minimizing the mean-squared TD loss using the Adam optimizer with gradient clipping to ensure numerical stability.

Actor update: In SAC, the actor maximizes the expected Q-value regularized by entropy, as given in Eq.19 [15]:

$$\nabla_\phi J(\phi) = \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi_\phi(\cdot|s)}\big[\nabla_\phi\big(\alpha \log \pi_\phi(a \mid s) - Q_\theta(s, a)\big)\big] \tag{17}$$

In TD3, the actor update is performed only once every K critic updates, as in Eq.20 [15]:

$$\nabla_\phi J(\phi) = \mathbb{E}_{s\sim\mathcal{D}}\big[\nabla_a Q_\theta(s, a)\big|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)\big], \tag{18}$$

where $\mathcal{D}$ denotes the replay buffer. It is worth noting that SAC updates the actor at every gradient step, whereas TD3 updates it less frequently (every K critic steps) to avoid destabilizing the critic.

979-8-3315-8138-1

Target network update: Both algorithms use soft target updates (Polyak averaging) to maintain stable target networks, as given in (19):

$$\begin{aligned} \theta_i' &\leftarrow \tau\,\theta_i + (1-\tau)\,\theta_i', \qquad i \in \{1,2\}, \\ \phi' &\leftarrow \tau\,\phi + (1-\tau)\,\phi'. \end{aligned} \tag{19}$$

III. Simualtion Results

A. Input data

The DRL algorithms are trained using a one-year time-series dataset that includes hourly EV load, PV generation, and electricity market prices. The EV charging profiles are constructed based on arrival and departure times, daily mileage, and specifications of commercially available EV models in Canada for the Edmonton region, following the methodology of [16]. The PV generation data for the same region are extracted from [17], and the corresponding TOU electricity prices are obtained from [18]. A representative weekday illustrating the hourly market prices (buying/selling), PV output, and EV load profile is shown in Fig. 2. The main BESS parameters used in training and simulation are summarized in Table I.

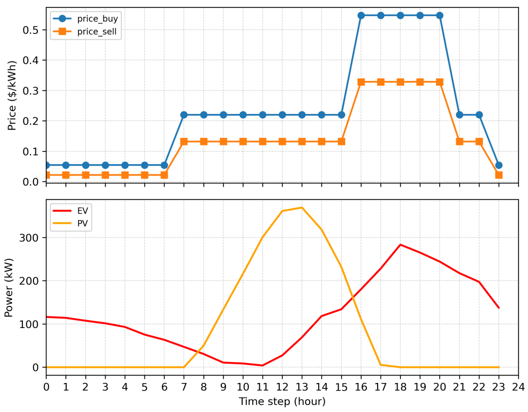

Fig. 2. Electricity prices, EV demand, PV power profiles for a selected day.

TABLE I
ESS parameters used in simulations.

Parameter	Value
Battery capacity C	500 kWh
Charging / Discharging efficiency η	0.98
Ramp rate limit ΔP_{ramp}	50 kW

B. Training and Testing Process

During training, the DRL agents interact with the simulated charging station environment over multiple episodes, using the one-year time-series dataset to learn optimal control policies through trial and error. The agents update their actor–critic networks based on observed rewards and system responses until convergence. After training, the learned policies are evaluated on unseen days to verify their generalization capability under varying PV, EV load, and price conditions. The convergence behaviour of both DRL methods is illustrated in Figure 3, which shows the moving-average episode returns (window = 200). The total training time was approximately 180 minutes for SAC over 18,000 episodes, and 420 minutes for TD3 over 25,000 episodes, demonstrating the differing convergence characteristics of the two algorithms. MILP was solved using CPLEX v22.1.2 with a run time of about 1-2 minutes.

C. Performance Comparison

The hourly power balance obtained using the SAC and TD3 is shown in Fig. 4 for a representative winter weekday (day 52). Both methods exhibit similar operational trends. During off-peak hours, the EV charging demand is primarily supplied by energy purchased from the grid while the ESS simultaneously charges. During peak demand periods, the ESS discharges to support EV load, and any excess PV generation (particularly during high-price intervals) is exported to the grid. These behaviors are further illustrated in Fig. 5, where both SAC and TD3 charge the ESS early in the morning and discharge it during evening peak hours. The DRL results are compared with a benchmark MILP solution (a standard scheduling formulation [19]), showing that SAC and TD3 remain close to the mathematical optimum, with slight deviations for SAC and more noticeable ones for TD3.

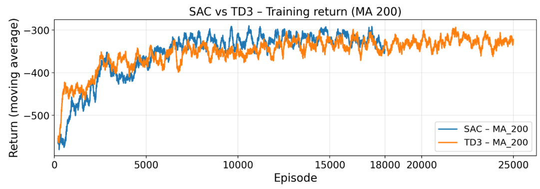

Fig. 3. Training returns for SAC and TD3 algorithms.

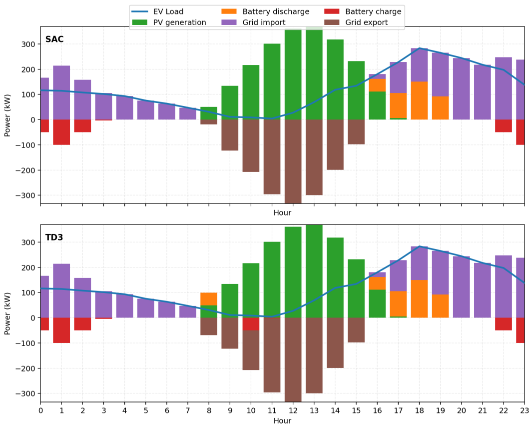

Fig. 4. Power balance in the test system during a 24-hour horizon.

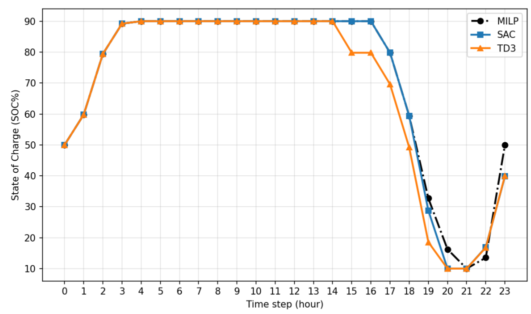

Fig. 5. SOC comparison among MILP, SAC, and TD3 for the selected day.

D. Analysis for Different Day Types

In this section, the performance of the trained models is evaluated under different load conditions, including both weekdays and holidays. Figure 6 compares the daily SOC behavior obtained from SAC and TD3 for two representative days: a

979-8-3315-8138-1

typical weekday (day 300) and a weekend day (day 168). Both algorithms exhibit similar operational patterns (charging the ESS during early morning low-price hours, discharging during the evening peak, and restoring the SOC to its initial level by the end of the day) demonstrating consistent and reliable behavior across varying demand conditions. However, slight differences can be observed during the weekday. Table II summarizes the annual and daily operating costs achieved by the three methods. As expected, the MILP reference model yields the lowest cost since it represents the mathematical optimum. The SAC algorithm performs very close to this benchmark, with only a 0.6% deviation from the optimal value. While TD3 follows similar operational trends to SAC and MILP, its total cost remains slightly higher, reflecting less efficient performance. This gap is attributed to the higher complexity and sensitivity of TD3's hyperparameter tuning, making convergence to near-optimal policies challenging.

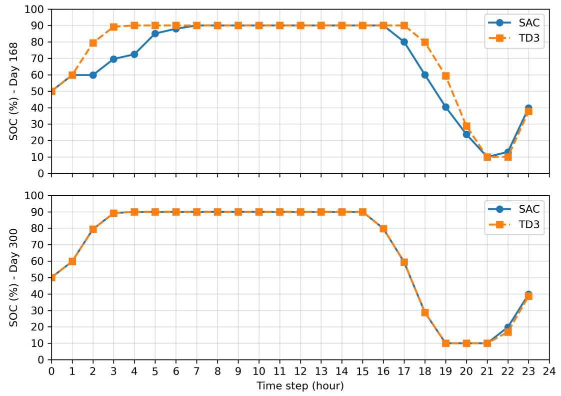

Fig. 6. SOC comparison among SAC, and TD3 methods for different days.

TABLE II
NET ANNUAL AND DAILY COST COMPARISON.

Method	Net annual cost ($)	Daily avg ($/day)
MILP	68,588	187.91
SAC	72,520	198.69
TD3	80, 756	221.25

IV. Conclusion

In this study, a deep reinforcement learning–based energy management framework was developed for a residential charging station equipped with PV and an ESS. The proposed methods were benchmarked against an MILP optimization model. Results show that both SAC and TD3 accurately capture the dynamic behaviour of the optimal scheduling strategy and closely match the MILP reference in terms of SOC evolution and net operating cost. Small deviations were noted for TD3, largely due to the algorithm's higher sensitivity to hyperparameter tuning, which can make convergence toward near-optimal policies more challenging.

Based on the findings of this study, future research can be carried out for integration of multi-site EV charging station integration via an aggregator coordinating each site's net grid exchange. Since the controller requires only a few key signals (such as PV, battery SOC, prices, and EV demand), the communication overhead is small and in the presence of latency, holding the last feasible setpoint can be regarded as a safe fallback.

Acknowledgments

This research has been supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada grant number RGPIN-2024-04197.

References

[1] I. Mahmud, M. B. Medha, and M. Hasanuzzaman, "Global challenges of electric vehicle charging systems and its future prospects: A review," *Research in Transportation Business Management*, vol. 49, p. 101011, 2023.

[2] S.-K. Hong, S. G. Lee, and M. Kim, "Assessment and mitigation of electric vehicle charging demand impact to transformer aging for an apartment complex," *Energies*, vol. 13, no. 10, 2020.

[3] S. Lokhande, Y. Bichpuriya, and V. Sarangan, "Real-time management of deviations in the demand of electric vehicle charging stations by utilizing ev flexibility," *Journal of Energy Storage*, vol. 97, p. 112719, 2024.

[4] G. S. Jagannath and R. B. Sharma, "Enhancing demand-side management in smart grids with electric vehicle and renewable energy integration," *Journal of Energy Storage*, vol. 140, p. 118985, 2025.

[5] N. R. Asr and M.-Y. Chow, "Network coordinated distributed demand management for optimal large-scale charging of phevs/pevs," in *2013 IEEE Power Energy Society General Meeting*, 2013, pp. 1–5.

[6] X. Cao, S. Striani, J. Engelhardt, C. Ziras, and M. Marinelli, "A semi-distributed charging strategy for electric vehicle clusters," *Energy Reports*, vol. 9, pp. 362–367, 2023, the 8th International Conference on Sustainable and Renewable Energy Engineering.

[7] P. Richardson, D. Flynn, and A. Keane, "Local versus centralized charging strategies for electric vehicles in low voltage distribution systems," *IEEE Transactions on Smart Grid*, vol. 3, no. 2, pp. 1020–1028, 2012.

[8] L. Wang, A. Dubey, A. H. Gebremedhin, A. K. Srivastava, and N. Schulz, "Mpc-based decentralized voltage control in power distribution systems with ev and pv coordination," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2908–2919, 2022.

[9] I. A. Zenhom, M. F. Shaaban, and W. A. Omran, "Grid interactive charging of evs in pv-powered parking lots considering uncertainties," *IEEE Access*, vol. 11, pp. 111 292–111 301, 2023.

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

[11] J. Zhang, J. Yan, Y. Liu, H. Zhang, and G. Lv, "Daily electric vehicle charging load profiles considering demographics of vehicle users," *Applied Energy*, vol. 274, p. 115063, 2020.

[12] H. Liang, Z. Wang, G. Hao, L. Xiao, and Z. Sun, "Deep reinforcement learning for charging electric vehicles with uncertain photovoltaic power," in *2022 China Automation Congress (CAC)*, 2022, pp. 1225–1231.

[13] F. Sun, S. Su, R. Diao, H. Cheng, D. Meng, and S. Lu, "Prediction-based ev-pv coordination strategy for charging stations using reinforcement learning," in *2022 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*, 2022, pp. 1–6.

[14] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 1861–1870.

[15] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 1587–1596.

[16] A. Hussain and P. Musilek, "Utility-scale energy storage system for load management under high penetration of electric vehicles: A marginal capacity value-based sizing approach," *Journal of Energy Storage*, vol. 56, p. 105922, 2022.

[17] "Renewables.ninja," https://www.renewables.ninja, accessed: 15 Nov. 2025.

[18] EPCOR Utilities Inc., "Residential electricity rates – collingwood and area," https://www.epcor.com/ca/en/on/collingwood-area/account/rates/residential-rates.html, accessed: 15 Nov. 2025.

[19] V.-H. Bui, A. Hussain, and H.-M. Kim, "Double deep *q*-learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 457–469, 2019.

979-8-3315-8138-1