CHALMERS
UNIVERSITY OF TECHNOLOGY

Smart Electric Vehicle Charging Algorithm to Reduce the Impact on Power Grids: a Reinforcement Learning Based Methodology

Citation for the original published paper (version of record):
Rossi, F., Diaz-Londono, C., Li, Y. et al (2025). Smart Electric Vehicle Charging Algorithm to Reduce the Impact on Power Grids: a Reinforcement
Learning Based Methodology. IEEE Open Journal of Vehicular Technology, In Press.
http://dx.doi.org/10.1109/OJVT.2025.3559237

Smart Electric Vehicle Charging Algorithm to Reduce the Impact on Power Grids: a Reinforcement Learning Based Methodology

Federico Rossi, *Graduate Member, IEEE,* Cesar Diaz-Londono, *Senior Member, IEEE,* Yang Li, *Senior Member, IEEE,* Changfu Zou, *Senior Member, IEEE,* and Giambattista Gruosso, *Senior Member, IEEE*

***Abstract*—The increasing penetration of electric vehicles (EVs) presents a significant challenge for power grid management, particularly in maintaining network stability and optimizing energy costs. Existing model predictive control (MPC)-based approaches for EV charging and discharging scheduling often struggle to balance computational efficiency with real-time operationability. This gap highlights the need for more advanced methods that can effectively mitigate the impact of EV activities on power grids without oversimplifying system dynamics. Here, we propose a novel scheduling methodology using a pre-trained Reinforcement Learning (RL) framework to address this challenge. The method integrates real grid simulations to monitor critical electrical points and variables while simplifying analysis by excluding the influence of real grid dynamics. The proposed approach formulates the scheduling problem to minimize costs, maximize rewards from ancillary service delivery, and mitigate network overloads at specified grid nodes. The methodology is validated on a benchmark electric grid, where realistic charging station utilization scenarios are simulated. The results demonstrate the method's robustness and ability to efficiently cope with the EV smart scheduling problem.**

***Index Terms*—Electrical Vehicle Scheduling, V2G, Reinforcement learning**

I. Introduction

Electric vehicles (EVs) have rapidly evolved from a niche market to a global phenomenon. With the global EV stock reaching 40.5 million in 2023 and 16% of new car sales in Europe in 2024, their adoption is reshaping the future of transportation [1]. While this transition supports a more sustainable future, it also presents significant challenges for power grids, particularly at the distribution level.

The widespread adoption of EVs leads to voltage fluctuations, grid imbalances, harmonic distortion, and increased peak hour loads, ultimately straining equipment and deteriorating power quality [2]. In particular, the uncoordinated charging of large numbers of EV overloads transformers and cables, shortening their lifespan and necessitating costly infrastructure upgrades [3], [4]. By 2050, projections indicate that four out of seven transformers could exceed their operational capacity due to rising energy demand from EV charging [5].

Smart charging strategies and Vehicle-to-Grid (V2G) technology are crucial in mitigating these effects and ensuring efficient energy management. In this context the optimization and scheduling of EVs charging activities is one of the most significant challenges that must be addressed to limit their impact on the electric distribution network. Various approaches have been proposed to address this challenge, including centralized and decentralized solutions [6]. For example, [7] analyzes the impact of growing EV charging demand on grid stability, while [8] examines parking lot-based charging strategies and their impact on cost and voltage stability. In [9] the authors highlights smart charging solutions to reduce grid costs and improve renewable integration. On the other side, to manage uncertainties associated with large-scale deployment of EVs and photovoltaic, [10] proposes a peer-to-peer energy trading framework, ensuring grid security and cost-effectiveness. Finally, [11] introduces a two-stage methodology to optimize charging costs and guarantee grid stability.

Among the various methods analyzed for addressing EV charging scheduling, Reinforcement Learning (RL) and Model Predictive Control (MPC) are two of the most commonly used approaches. A systematic review of RL applications in charging scheduling is provided in [12], highlighting key algorithms, challenges, and future research directions. For instance, to minimize charging time in public stations, [13] formulates the scheduling problem as a Markov decision process (MDP) and applies deep RL, showing significant improvements over baseline methods. The impact of RL on optimizing photovoltaic self-consumption and EV state of charge is examined in [14], where it is compared with the rules-based and MPC strategies, demonstrating its potential to improve efficiency in managing energy use. Similarly, [15] proposes a multi-agent RL approach for optimal EV charging and discharging scheduling, ensuring grid stability while minimizing costs. Lastly, [16] reviews RL-based frameworks for EV energy management, analyzing coordination strategies and optimization techniques for charging under uncertainty. MPC-based approaches also play a crucial role in optimizing EV charging. In [17], a MPC approach is proposed to manage demand charges in real-time EV charging scheduling, showing improvements in operational profit and charging scheduling

F. Rossi, and G. Gruosso are with Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy (e-mail: federico1.rossi@polimi.it, giambattista.gruosso@polimi.it)

C. Diaz-Londono is with Microgrid and Renewable Energy Research Center, Huanjiang Laboratory, Zhejiang University, Zhuji, China (e-mail: cesar.diaz@ieee.org)

Y. Li and C. Zou are with Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden (e-mail: yang.li@chalmers.se, changfu.zou@chalmers.se)

under demand charge constraints. In [18] the authors introduce a smart EV charging pool algorithm using optimal control, aiming to minimize operational costs while ensuring flexibility in charger management and reducing power demand. In [19], two smart charging coordinators are proposed to manage EV charging, optimizing grid integration and preventing transformer overloads across various scenarios.

Under a more technical point of view, RL learns optimal policies by interacting with the environment, making it highly adaptable to dynamic and uncertain conditions. While computationally intensive during the training phase, it can deliver superior performance once trained, particularly in handling complex, nonlinear systems. Conversely, MPC relies on an accurate mathematical model of the system to predict future states and optimize control actions over a finite horizon. It excels at managing multi-input, multi-output (MIMO) systems and handling constraints. MPC can be more robust than RL in scenarios where the system model is well defined and the uncertainties are minimal. However, its performance is limited by the need for accurate system modeling, which is challenging to achieve in highly dynamic and uncertain environments, such as EV charging networks. Moreover, MPC can be computationally demanding, especially when applied to large-scale systems or long prediction horizons, potentially limiting its feasibility for real-time applications. On the other hand, RL shows significant potential for optimizing EV scheduling activities [20]. Indeed, RL is particularly effective in dynamic scheduling environments where conditions change frequently. This adaptability is essential for EV scheduling, as it must account for variations in energy demand, charging station availability, and traffic conditions. In addition, RL can make real-time decisions by managing uncertainty and environmental fluctuations, which are common in EV scheduling due to unpredictable energy consumption patterns and the variable availability of renewable energy sources. These peculiarities are detailed in several works, such as in [21], where the authors explore RL for coordinating EV charging with grid services, considering battery aging and user satisfaction. Additionally, [22] proposes a deep RL approach to optimize charging costs and reduce waiting times. In [23], deep Q-learning is applied to improve large-scale charging scheduling during peak demand periods, while [24] integrates RL with voltage control to enhance distribution network stability.

In light of the conducted literature review, we identified key areas that require further investigation. First, we observed that many works solve the optimal EV scheduling problem relying on MPC, a method that, while effective, has computational limitations in real-time applications. Second, several studies, such as [14] and [22], simplify the grid model to mere power balances, overlooking the impact of real grid dynamics on the results. Furthermore, most RL applications lack pretraining, resulting in slower convergence and suboptimal performance compared to MPC and standard optimization techniques.

With ACER [25] promoting V2G technologies at the European level, several countries have begun initial testing phases. In this context, this work focuses on developing and validating technical solutions to integrate V2G into the energy system. This is achieved by creating a replicable model for network simulations, intended for users seeking to optimize vehicle management and improve overall grid efficiency in real-world applications. In particular, this paper presents a scheduling problem geared toward real-time applications and overcomes the traditional limitations of MPC-based methods. Furthermore, the primary objective is to mitigate the impact on power grids, so the proposed methodology integrates simulation of the real grid to monitor its critical electrical points and variables neglecting the impact of real grid dynamics on the results. In order to accelerate the learning process, it may be appropriate to rely on forms of pretraining that can accelerate convergence and make the problem more robust. The work therefore focuses on integrating these aspects and in particular aims to:

- Formulate the EVs charging and discharging scheduling problem as a MDP with the goal of minimizing costs, maximizing rewards from the V2G service provision, and limiting grid overloads. This approach leverages state-of-the-art Proximal Policy Optimization (PPO) algorithms to improve learning efficiency and accuracy
- Improving performance by using Neural Network (NN) pretraining techniques based on expert-generated trajectories derived offline.
- Integrating power flow simulation of a power grid where scheduling is applied. This algorithm incorporates power flow calculations to evaluate transformer load levels and overall network state, ensuring that network dynamics are accurately accounted for. In addition, the algorithm takes into account the actual characteristics of EV batteries to optimize charging and discharging decisions.
- Employ the model-inversion method and a 0th-order equivalent circuit model for a precise calculation of the operational range of EV batteries.

The remainder of the paper is organized as follows. Section 2 describes the problem formulation and provides an overview of RL. Section 3 presents the model used to run the simulation. Section 4 describes the RL-based model used to solve the optimization problem. Results are reported in Section 5 and discussion, while Section 6 concludes the paper with suggestions for future work.

II. Problem Formulation and Preliminaries

A. Problem Formulation

This work aims to develop an aggregator model for managing EVs charging within a distribution network, optimizing charging/discharging processes to minimize end-user costs and reduce network overloads. A medium-voltage CIGRE network is employed to simulate the distribution network, containing distributed generation units such as photovoltaic and wind systems, various nodes, and transformers between MV and LV levels. The original network model is detailed in [26], and its specific implementation is described in [27]. Stochastic loads, derived from real load profiles, are applied to all network nodes, while a finite number of charging stations (each with a variable number of charging points) is assigned only to selected nodes. The management of charging and discharging

cycles for V2G, is governed by an optimization algorithm based on Deep Reinforcement Learning (DRL).

The implemented model follows an iterative operational flow over discrete time intervals $t \in \mathcal{T}$. In each time step t, the model operates as follows: 1) Initially, the network state is determined through a Power Flow (PF) calculation based on the loads and distributed generators present. This provides the current system conditions in terms of voltages, overloads, and node capacities. Based on these results, the maximum power deliverable to each Charging Station (CS) $i \in \mathcal{I}$ is calculated, taking into account transformer capacity limitations and general system conditions. This value serves as a key input parameter for the next step, where energy flows for the CSs are computed by the DRL algorithm. 2) The optimization of charging and discharging power for each EV $k \in \mathcal{K}$ is based on a RL model that takes as input forecast data, the vehicle presence vector, and the maximum power deliverable by each station. Three charging stations are considered, each equipped with six AC charging points. Within the simulation time window, multiple vehicles can sequentially undergo charging and discharging processes. Each vehicle State of Charge (SoC) and departure time are continuously monitored. At any given time, all six charging points at a station can be occupied or free, determined by a random algorithm. The EVs are modeled with diverse real-world characteristics, reflecting different types and capabilities. Using the PPO algorithm, the model determines the optimal charging/discharging profile for each EV in each time interval, minimizing costs while staying within the power limit derived in the previous step to avoid network overloads. 3) After defining the charging profile, the total power of each station is reintegrated into the network, performing a new PF calculation to update the network state, considering the contribution of the charging points. In this way, the impact of charging on the network is evaluated, and voltage variations and overload conditions are assessed.

In the next subsections are described the basis of RL and the PPO, i.e. the algorithm used to control the power exchanged among the EVs and the grid.

B. Principles of RL

RL is a machine learning (ML) paradigm where an agent interacts with an environment to maximize a long-term reward. This interaction is modeled as a MDP, represented by the tuple $(\mathcal{S}, \mathcal{A}, P, R)$, where:

- $\mathcal{S}$ is the set of states;
- $\mathcal{A}$ is the set of actions;
- $P(s_{t+1} \mid s_t, a_t)$ is the transition probability to the next state given the current state and action;
- $R(s, a, r)$ is the reward function, providing the immediate reward r for a given state-action pair;

At each time step t, the agent observes a state s_t, selects an action a_t based on the policy receives a reward r_t, and transitions to the next state s_{t+1}. There are two main types of policies, deterministic and stochastic ones, expressed by $\mu(a_t|s_t)$ and $\pi(a_t|s_t)$ respectively. The goal of RL is to find an optimal policy π_* that maximizes the expected cumulative reward, or return, defined as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (1)$$

RL methods are typically divided into three categories: value-based, policy-based, and actor-critic methods. Value-based methods, such as Q-learning, aim to compute the optimal Q-values iteratively using the Bellman equation and then derive the optimal policy. While value-based methods are effective in discrete action spaces, they face limitations in continuous action environments. To overcome this, policy-based methods were developed. They leverage on the policy gradient method to find the optimal parameter θ of the NN to obtain the correct probability distribution over the action space. The goal is to assign high probabilities to the actions that maximize the cumulative reward of a trajectory τ.

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)}[G(\tau)] \quad (2)$$

The objective function is maximized by computing the gradient $\nabla_\theta J(\theta)$ and updating the parameters according to:

$$\theta = \theta + \alpha \nabla_\theta J(\theta) \quad (3)$$

Actor-critic methods combine both value-based and policy-based approaches. They use two networks: the actor-network, which is responsible for updating the policy, and the critic network, which evaluates the policy produced by the actor. The parameters of both networks are updated at each step of the episode. This combination helps stabilize training by benefiting from both the direct policy optimization of policy-based methods and the value estimation of value-based methods.

C. PPO Algorithm with Clipped Surrogate Loss

Policy gradient methods are on-policy algorithms, meaning they improve the same policy used to generate trajectories during each iteration. As previously shown in (3), these methods adjust the network parameters by computing the gradient of the objective function with respect to the policy parameters, often using a small learning rate α to avoid large deviations between the old and new policies. This approach helps mitigate issues such as model collapse, where the policy updates are too large, destabilizing training. However, taking small steps with a small learning rate can slow down learning. Trust Region Policy Optimization (TRPO) [28] addresses this issue by attempting to make larger policy updates while ensuring that the new policy does not differ too much from the old one. This is achieved by constraining the Kullback-Leibler (KL) divergence between the old and new policies to be below a threshold, δ. This constraint is known as the trust region constraint, which allows TRPO to make larger updates without risking instability. The downside is that TRPO is computationally expensive because it requires second-order optimization methods. PPO [29] improves upon TRPO by simplifying the process. PPO ensures that the policy updates stay within the trust region using a first-order method, which makes it computationally more efficient. Instead of using direct constraints in the objective function, PPO uses a clipping

function (in the PPO-Clip variant) to limit how much the new policy can differ from the old one. This ensures that the updates remain stable and efficient without needing complex second-order methods. There are two main variants of PPO: PPO-Penalty and PPO-Clip. The second variant is used in this paper, and its pseudocode is presented in Alg. 1.

Algorithm 1 PPO-Clip

1: Input: initial policy parameters θ_0, initial value function parameters ϕ_0
2: **for** $k = 0, 1, 2, ...$ **do**
3: Collect a set of trajectories $\mathcal{N}_k = \{\tau_i\}_{i=1}^{N}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
4: Compute rewards-to-go R_t for each state s_t.
5: Estimate the advantage $\hat{A}_t$, based on the current value function V_{ϕ_k}, using:

$$\hat{A}_t = Q(s_t, a_t) - V_{\phi_k}(s_t) \tag{4}$$

or, using Generalized Advantage Estimation (GAE).
6: Compute the gradient of the objective function $\nabla_\theta L_{\text{clip}}$ where:

$$L_{\text{clip}} = \frac{1}{|\mathcal{N}_k|} \sum_{\tau \in \mathcal{N}_k} \sum_{t=0}^{T-1} \min \Big(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \Big) \tag{5}$$

and:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} \tag{6}$$

7: Update the policy network parameter θ typically via gradient ascent with Adam:

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta L_{\text{clip}} \tag{7}$$

8: Compute the mean square error (MSE) of the value network

$$J(\phi) = \frac{1}{|\mathcal{N}_k|} \sum_{\tau \in \mathcal{N}_k} \sum_{t=0}^{T-1} \Big(V_\phi(s_t) - \hat{R}_t \Big)^2 \tag{8}$$

9: Compute the gradient of the value network $\nabla_\phi J_\phi$
10: Update the value network parameter ϕ typically via gradient descent with Adam:

$$\phi_{k+1} = \phi_k - \beta \nabla_\phi J_\phi \tag{9}$$

11: **end for**

III. Model framework

To implement the above strategy, a Python-based simulation framework has been developed. It enables the modeling of all aspects of EV creation and management, as well as the evaluation of their impact on the network. The following subsections outline its various components.

A. Electric Vehicles

In the proposed energy management model for EVs, a detailed database is used, including key specifications of various vehicle models such as *battery capacity*, *maximum charging and discharging power*, and support for V2G functionality. This data enables the construction of a management system that simulates the real behavior of batteries during charging and discharging cycles. The real-world characteristics of 12 different vehicles have been considered. The data can be found in [30], which is based on the RVO-NL report. The maximum and minimum power absorbable by each vehicle are defined by $\overline{P}^{\text{ch}}$ and $\underline{P}^{\text{ch}}$, respectively. Similarly, the power deliverable during discharge is limited by $\overline{P}^{\text{dis}}$ and $\underline{P}^{\text{dis}}$. These values are determined using an empirical methodology that maps the maximum power value based on the SoC of a vehicle at a given moment, according to its characteristics. These parameters vary depending on the charging mode, AC or DC. The derivation of this mapping is described below.

TABLE I
Coefficients for the Open Circuit Voltage (OCV) Polynomial and Battery Cell Specifications.

Parameter	Value
Open Circuit Voltage (OCV) Polynomial Coefficients	
γ_4	-3.4599
γ_3	8.0326
γ_2	-5.8485
γ_1	2.1021
γ_0	3.3324
Battery Cell Specifications	
Nominal capacity	$2.05\,\text{Ah}$
Nominal voltage	$3.6\,\text{V}$
Lower voltage limit	$2.5\,\text{V}$
Upper voltage limit	$4.2\,\text{V}$
Internal resistance	$0.035\,\Omega$

In this paper, the same *battery cell specifications* are considered for all battery packs of different vehicles. The parameters that vary are $\overline{P}^{\text{ch}}$, $\underline{P}^{\text{ch}}$, $\overline{P}^{\text{dis}}$, $\underline{P}^{\text{dis}}$, and the EV battery capacity E, to account for the actual power limitations defined by the Battery Management Systems (BMS) of each vehicle.

1) Equivalent Circuit Model for the Battery: The battery considered for the model is a Sanyo Li-ion 18650 cell, model UR18650E, characterized by an anode composed of Li(NiMnCo)O_2 and a carbon cathode. The main specifications of the cell are listed in Table I. To represent the electrical behavior of the battery, a 0th-order Equivalent Circuit Model (ECM) is used [31]. In this model, the battery voltage $V(t)$ is represented as:

$$V(t) = \text{OCV}(z(t)) + R \cdot I(t) \tag{10}$$

where:

- OCV$(z(t))$: Open Circuit Voltage (OCV), a function of the SoC $z(t)$.
- R: internal resistance of the battery.
- $I(t)$: current, the control variable of the battery. When positive, it indicates charging; when negative, it indicates discharging.

As illustrated in (11), the derivative of the SoC $z(t)$ can be expressed as the ratio between the current and the nominal ca-

pacity of the battery $b_{\text{cap,Ah}}$ in ampere-hours (Ah). Integrating this equation over time with a step size Δt, the SoC is updated iteratively for each time interval. Equation 12 enforces that the SoC $z(t)$ remains within the range of 0 (fully discharged) to 1 (fully charged):

$$\frac{dz(t)}{dt} = \frac{I(t)}{b_{\text{cap,Ah}}} \tag{11}$$

$$0 \leq z(t) \leq 1 \tag{12}$$

The OCV is modeled as a fourth-degree polynomial based on voltage measurements at different SoC levels. The coefficients obtained from the fit are shown in Table I. The resulting polynomial is expressed in (13).

$$\text{OCV}(t) = \gamma_4 \cdot z(t)^4 + \gamma_3 \cdot z(t)^3 + \gamma_2 \cdot z(t)^2 + \gamma_1 \cdot z(t) + \gamma_0 \tag{13}$$

2) Power Calculation: Once the OCV curve as a function of the SoC is obtained, the charge/discharge current is calculated using the model-inversion method [32]. Since both current and voltage must remain within safe limits, two current limits are defined:

- Maximum current limited by power, I_{cp}:

$$I_{\text{cp}} = \frac{-\text{OCV}(z(t)) + \sqrt{\text{OCV}(z(t))^2 + 4 \cdot P_{\text{max}} \cdot R}}{2R} \tag{14}$$

- Maximum current limited by voltage, I_{cv}:

$$I_{\text{cv}} = \frac{V_{\text{max}} - \text{OCV}(z(t))}{R} \tag{15}$$

The actual current is then constrained to the minimum value between I_{cp} and I_{cv} to ensure that neither the maximum power nor the maximum voltage are exceeded:

$$I(t) = \min(I_{\text{cp}}, I_{\text{cv}}) \tag{16}$$

This approach ensures that the battery always operates within safe limits, avoiding overcurrents and overvoltages. Once the current for a given time instant t is determined, the updated SoC can be derived as follows:

$$z(t+1) = z(t) + \frac{I(t)}{3600 \cdot Q} \cdot T_s \tag{17}$$

At each time step, the power $P(t)$ delivered or absorbed by the battery is calculated as the product of the battery voltage $V(t)$ and the current $I(t)$:

$$P(t) = V(t) \cdot I(t) \tag{18}$$

This method provides a power map as a function of the SoC as the one in Fig. 1, which is useful for evaluating battery behavior and the maximum deliverable/absorbable power under various charge/discharge conditions. The obtained values are calculated at the battery cell level and must therefore be scaled proportionally to the capacity of the vehicle under consideration to reach the battery pack level.

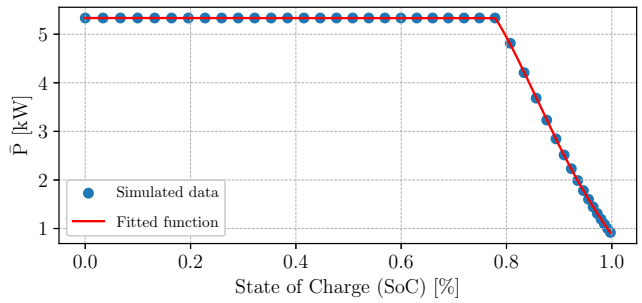


Fig. 1. SoC - $\overline{P}$ map for a battery of 11kWh.

B. Charging Station Simulation

In the simulator, the simultaneous charging of a group of EVs is centrally managed by a CS $i \in \mathcal{I}$. Each vehicle $k \in \mathcal{K}$, with $\mathcal{K}$ being the set of all vehicles, has a variable initial SoC and specific arrival and departure times. These values are randomly determined within the time window during which the simulation takes place. The station simulation dynamically monitors vehicle availability and their remaining energy, allocating charging power while respecting the operational limits of the station. The parameters used in the simulation are defined as follows:

- $\text{SOC}_k(t)$: current state of charge of vehicle k at time $t \in \mathcal{T}$.
- $\text{SOC}_k^{arr}(t)$: arrival state of charge of vehicle k.
- $\text{SOC}_k^{target}(t)$: minimum state of charge that vehicle k must have upon departure.
- t_k^{arr} and t_k^{dep}: arrival and departure times of vehicle k.
- $u_{k,t}$: binary variable indicating whether vehicle k is connected to the station at time t (1 if connected, 0 otherwise).
- N: total number of vehicles managed by the station.
- $t \in \mathcal{T}$: current time in the simulated day.

At each time step $t \in \mathcal{T}$, the vehicles present and those departing within the next hour are initially identified. Suppose a vehicle $k \in \mathcal{K}$ is connected and leaving in the immediately following time step. In that case, it is added to a list of vehicles leaving the station in the next interval. This allows for assessing the SoC of the departing vehicle to apply a potential penalty if its SoC is lower than the target value set. This logic is the same used in [33]. In addition to this, the station imposes constraints on the total available power to avoid overloading, with dynamic power distribution to the charging EVs:

- Station Maximum Power Constraint: the sum of the powers assigned to the vehicles must respect the maximum station capacity $\overline{P}_i^{cs}$ at all times $t \in \mathcal{T}$.
- Power Constraint for each EV: the power allocated to each EV cannot exceed the individual limits.

This mechanism ensures that the station's total power usage remains within the maximum allowable limit, preventing overloads while maintaining operational stability.

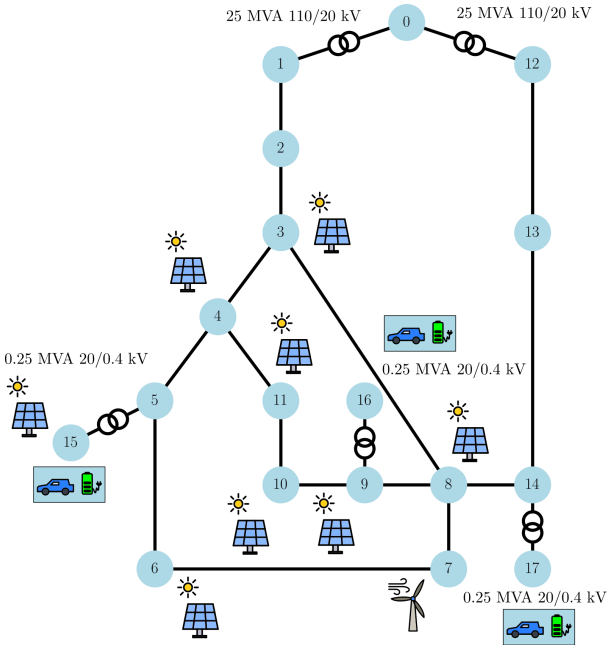


Fig. 2. CIGRE distribution network. Buses 15, 16, and 17 are equipped with charging stations. The diagram also allows to identify the types of generators installed at each bus, where present. Additionally, the two types of transformers used in the system are shown.

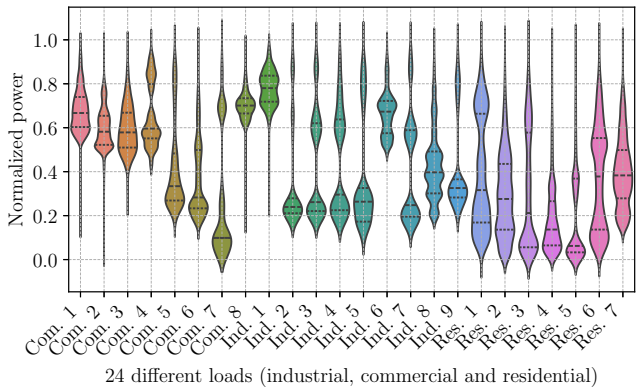


Fig. 3. Distribution of active power for the three main load categories: industrial, commercial, and residential. Each category includes multiple load profiles, which have been adjusted and randomized to simulate real scenarios.

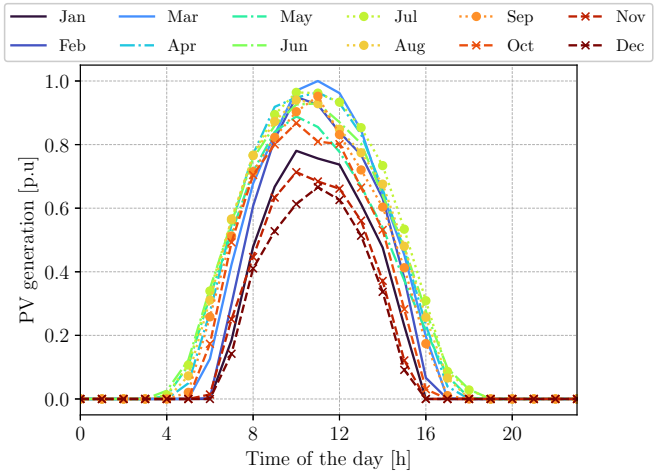


Fig. 4. Hourly generation profiles for PV generators, calculated based on nominal capacity and expected generation values.

C. Electric Grid

An important feature of the simulator used is its capability to include a physical model of the grid and perform PF calculations. As previously mentioned, the simulated electrical grid is based on the CIGRE MV network depicted in Fig. 2. The model was constructed using the `pandapower` package [34]. It includes a set of MV buses $b \in \mathcal{B}$ interconnected by distribution lines, loads $l \in \mathcal{L}$, distributed generators $g \in \mathcal{G}$ (considering both photovoltaic (PV) and wind sources), transformers $t \in \mathcal{T}$, and various switches $s \in \mathcal{S}$. Additionally, to allow for the inclusion of a fixed number of CSs $c \in \mathcal{C}$, three low-voltage LV buses were added, interfaced to the grid via three transformers. These transformers are of type `0.25 MVA 20/0.4 kV`, suitably rated to support loads up to 0.25 MVA.

The active power demand of the loads is obtained from a real-world dataset [35], while the reactive power (Q) is calculated using a power factor of $\cos(\varphi) = 0.95$. The distribution of active power for the normalized loads is illustrated in Fig. 3. As illustrated, the loads are grouped into three main categories: industrial, commercial, and residential. Multiple load profiles are defined for each category. These profiles were adjusted and randomized to simulate realistic variations and reflect the characteristics of loads connected to both MV and LV buses. Specifically, industrial and commercial loads are connected to MV nodes, while residential loads, along with CSs, are assigned to LV nodes. For PV generators, the power of each generator is derived by multiplying the nominal plant capacity by the expected generation value for a specific hour on a given day. These values, shown in Fig. 4, were obtained using a methodology similar to that described in [36] and the data collected by ENTSO-e [37]. Power flow calculations are performed at each time step t to assess the state of the network, calculate power flows along each branch, and monitor transformer loading. Transformer loads are continuously monitored to ensure they do not exceed 100%, thus avoiding rapid degradation caused by thermal effects from overloading. The loading percentage of each transformer (L_{trafo}) is calculated as:

$$\text{loading}_{\%} = \max\left(\frac{i_{\text{hv}} \cdot v_{\text{hv}}}{s_{\text{n\_mva}}}, \frac{i_{\text{lv}} \cdot v_{\text{lv}}}{s_{\text{n\_mva}}}\right) \cdot 100 \qquad (19)$$

where i_{hv} and i_{lv} are the currents on the high-voltage and low-voltage sides of the transformer, v_{hv} and v_{lv} are the voltages on the high-voltage and low-voltage sides, and $s_{\text{n\_mva}}$ is the transformer's nominal power in MVA.

To prevent overloading, the maximum additional power that can be drawn from the LV buses by individual CSs without exceeding the transformer's capacity is calculated as:

$$P_{\text{max LV}} = \frac{S_{\text{nom}} \cdot (100 - L_{\text{trafo}})}{100} \cdot \cos(\varphi) \qquad (20)$$

where $P_{\text{max LV}}$ represents the maximum power that can be safely added without compromising transformer performance.

IV. Proposed DRL-based charging strategy

This work focuses on maximizing the profit derived from providing V2G services. Specifically, the goal is to minimize charging and maximize discharging during periods of high energy demand, and vice versa, while satisfying user requirements, such as ensuring the vehicle is charged to the desired SoC upon departure, and respecting grid constraints, such as avoiding transformer overloads. In this scenario, the t^{arr} and t^{dep}, as well as their target SoC, are considered known. Additionally, the capacity E of the vehicle battery is assumed to be provided when connecting to the CS. Time is discretized into intervals $t \in \mathcal{T}$, and the number of vehicles $k \in \mathcal{K}$ varies dynamically. This optimization problem is solved for a single CS, as randomizing variables such as t_k^{arr}, t_k^{dep}, and $u_{k,t}$ enables the training of a NN capable of managing a generic station. Therefore, a single controller model can be developed and applied locally at each CS $c \in \mathcal{C}$. The optimization problem is formulated as follows.

Objective function:

$$\max_{P^{\text{ch}}, P^{\text{dis}}} \sum_{t\in\mathcal{T}} \sum_{k\in\mathcal{K}} \left(P_{k,t}^{\text{dis}} \cdot c_t^{\text{dis}} + P_{k,t}^{\text{ch}} \cdot c_t^{\text{ch}} \right) \cdot \Delta t \tag{21}$$

Subject to:

$$P_{k,t} = P_{k,t}^{\text{ch}} - P_{k,t}^{\text{dis}} \qquad \forall k, \forall t \tag{22}$$

$$\text{SOC}_{k,t+1} = \text{SOC}_{k,t} + \frac{P_{k,t}}{E_k} \cdot \Delta t \qquad \forall k, \forall t \tag{23}$$

$$\text{SOC}_{k,t=t_k^{\text{arr}}} = \text{SOC}_k^{\text{arr}} \qquad \forall k \tag{24}$$

$$\text{SOC}_{k,t=t_k^{\text{dep}}} \geq \text{SOC}_k^{\text{target}} \qquad \forall k \tag{25}$$

$$\underline{\text{SOC}}_k \leq \text{SOC}_{k,t} \leq \overline{\text{SOC}}_k \qquad \forall k, \forall t \tag{26}$$

$$\underline{P}_k^{ch} \leq P_{k,t}^{ch} \leq \overline{P}_k^{ch} \qquad \forall k, \forall t \tag{27}$$

$$\underline{P}_k^{dis} \leq P_{k,t}^{dis} \leq \overline{P}_k^{dis} \qquad \forall k, \forall t \tag{28}$$

$$\sum_{k\in\mathcal{K}} P_{k,t} \leq \overline{P}^{\text{cs}} \qquad \forall k, \forall t \tag{29}$$

$$P_{k,t}^{\text{ch}} \cdot P_{k,t}^{\text{dis}} = 0 \qquad \forall k, \forall t \tag{30}$$

In (21) $P_{k,t}^{\text{dis}}$ and $P_{k,t}^{\text{ch}}$ represent the discharging and charging power of vehicle k at time t, respectively, and c_t^{dis} and c_t^{ch} are the electricity prices in the day-ahead market for discharging and charging, respectively, at time t. The profit is calculated over the time step Δt. The difference between charging and discharging power is expressed as $P_{k,t}$ as shown in (22).

The second constraint (23) governs the evolution of SoC of each vehicle over time. The SOC at time $t+1$ depends on the previous SOC, adjusted for the charging and discharging powers. The third and fourth constraints, (24) and (25), set the SoC at arrival and departure times. Specifically, the SoC at arrival must match the initial value $\text{SOC}_k^{\text{arr}}$ (given as a random input), while the SoC at departure is constrained to the target value $\text{SOC}_k^{\text{target}}$. The constraint (26) ensures that the SOC of each vehicle remains within allowable bounds, $\underline{\text{SOC}}_k$ and $\overline{\text{SOC}}_k$, throughout the time period. The sixth and seventh constraints, (27) and (28), limit the charging and discharging power of each vehicle. Thus the power must remain within predefined bounds, which account for the physical limits of the vehicle's battery and charging system. The constraint (29) ensures that the total power requested from all vehicles at any given time does not exceed the power capacity $\overline{P}^{\text{cs}}$ of the CS. This is necessary to avoid overloading of the transformer. Lastly, the ninth constraint (30) ensures that a EV cannot charge and discharge simultaneously. This constraint is enforced by the product of the charging and discharging powers, which must be zero at all times, ensuring that the vehicle operates in a mutually exclusive mode for charging or discharging. The simplest, but also very effective, way to deal with these constraints is to consider them as soft constraints via loss functions [38].

A. Optimization Solution

In the proposed DRL-based framework for EV charging optimization the PPO agent is initially trained offline in a simulation environment that emulates the dynamic behavior of the distribution network and EV charging demands. Through this process, the agent learns an optimal policy π that determines the charging and discharging actions for each vehicle k at every time step t. The policy is subsequently applied in real-time to manage CSs efficiently and profitably, ensuring efficient power allocation while adhering to grid constraints and user requirements.

B. State and Action Spaces

The state, serving as the input to the PPO agent, encapsulates key information about the EV charging environment. This includes the maximum power deliverable by each CS, the actual electricity price, forecasts on future electricity prices for a time horizon $h = 24$ hours, SoC of each vehicle (if present), and the time remaining until each vehicle's departure (assumed to be known). All these features are normalized using the Min-Max normalization, thus remaining within the range [0, 1] scaling technique to handle the diverse ranges of the parameters before being processed by the agent:

$$S_t = \left[\overline{P}_i^{cs}, p_t, \tilde{p}_{t+1}, \tilde{p}_{t+h}, SoC_{1:k,i,t}, t_{1:k,i}^{dep} - t \right] \forall k, \forall i, \forall t \tag{31}$$

The action space consists of the charging and discharging power allocated to each vehicle:

$$a_t = \left[\Delta P_{1,i,t}, \ldots, \Delta P_{k,i,t} \right] \forall k, \forall i, \forall t \tag{32}$$

Differently from the state space, the action space is bounded between [-1, 1]. This is justified by the fact that EVs provide V2G services. This means that they can both absorb and withdraw energy. By interacting with the environment during training, the PPO agent learns to make incremental adjustments to charging decisions. This enables the agent, once trained, to dynamically optimize the charging process in real-time, achieving efficient energy utilization and compliance with user and system constraints.

C. Neural Network Structure

The NN used in the PPO framework consists of an actor-critic architecture, with both the actor and critic networks having identical structures. The input to the network corresponds to the state space defined in the previous section. The output of the actor-network represents the charging/discharging power allocated to each vehicle.

The model is composed of five fully connected layers with 400, 300, 128, 64, and k neurons, respectively, where k corresponds to the dimension of the action space, i.e. the number of EVs in the system. The first four layers employ the `tanh` activation function, which is well-suited for capturing non-linear relationships and preventing the vanishing gradient problem. The final output layer also applies the `tanh` function to ensure the outputs remain within a bounded range, reflecting the predicted charging power. This is necessary in order to comply with the action space constraints, which are bounded in the range of [-1, 1]. Indeed this is the optimal range for PPO when implemented with Stable Baselines 3 (SB3) [39]. This structure is used both for the actor and the critic in the PPO framework, ensuring that the agent can effectively optimize the charging process while adhering to the system's constraints.

D. Imitation Learning-Based Initialization

To improve the efficiency of the DRL training process for optimizing EV charging and discharging, Imitation Learning (IL) was employed to initialize the weights of the actor NN. Directly training the DRL agent from scratch in environments with large state and action spaces can be computationally expensive and may even fail to converge in some scenarios. The adoption of IL mitigates these challenges by leveraging expert demonstrations to provide a strong initialization for the actor NN, thereby reducing the number of steps required for policy optimization.

In this work, offline optimization was used to generate expert trajectories, where each trajectory maps network states s_t to optimal charging and discharging actions for EVs. These (state, action) pairs were then used to train the actor NN through supervised learning, effectively formulating the problem as a regression task. The training objective minimizes the mean squared error (MSE) between the predicted actions and the expert actions, as stated in (33). The actor NN was trained using the Adam optimizer with a batch size of 256 and a learning rate of 0.001.

$$\mathcal{L}(\theta) = \frac{1}{N_{\mathrm{IL}}} \sum_{(s_t,a_t)\in\mathcal{D}_{\mathrm{train}}} \left\| a_t - \mu_\theta(s_t) \right\|^2 \tag{33}$$

where $\mathcal{D}_{\mathrm{train}}$ is the training dataset containing N_{IL} samples, s_t represents the system state, a_t denotes the expert action, and μ_θ is the parameterized mapping function of the actor NN.

The trained weights serve as the initial parameters for the actor NN within the PPO framework. This initialization not only accelerates the convergence of the DRL agent but also provides a validation mechanism for the NN structure, ensuring its ability to capture relevant patterns in the state-action mapping. However, as IL alone may not generalize across all operating conditions of the network, the DRL process is subsequently employed to further refine the control policy through iterative interaction with the environment.

E. Offline Training Process of PPO Agent for effective Charging planning

The training of the PPO agent, as shown in Fig. 5, involves interaction with the environment in a series of episodes. Each episode starts with the `reset(·)` function, which initializes the system, retrieves the necessary data (e.g. vehicle specifications, energy prices, and grid data), and sets the initial state s_t. The agent then takes actions based on its current policy, which are passed to the `step(·)` function. This function computes the updated state, the resulting reward, and the `done` signal based on the agent's actions, and progresses to the next time step. The episode terminates when the last time step $t \in \mathcal{T}$ is reached, marking the end of the episode.

As previously mentioned, this paper adopts the soft constraints approach. This means that the previously described constrained optimization problem is converted into an unconstrained optimization problem by introducing penalty terms with fixed coefficients. Thus, the resulting reward function becomes:

$$Reward = -\left[R_{\mathrm{ec}} + R_{\mathrm{soc}} + R_{\mathrm{ol}}\right] \tag{34}$$

where ① R_{ec} is the general cost associated with the difference between the energy drawn from and injected into the grid ② R_{soc} is the penalty for charging incompleteness, applied when an EV SoC at departure is below the target value. The penalty increases as the final SoC deviates from this target ③ R_{ol} is the penalty for grid overload, applied when the total energy used by the EVs exceeds the grid's available power capacity at any given hour.

$$\begin{cases} R_{\mathrm{ec}} = \sum_{t\in\mathcal{T}} \sum_{k\in\mathcal{K}} \left(-P^{\mathrm{dis}}_{k,t} \cdot c^{\mathrm{dis}}_t + P^{\mathrm{ch}}_{k,t} \cdot c^{\mathrm{ch}}_t \right) \cdot \Delta t \\ R_{\mathrm{soc}} = \sum_{t\in\mathcal{T}} \sum_{k\in\mathcal{K}} \left(100 \cdot \left| \mathrm{SOC}^{\mathrm{target}}_k - \mathrm{SOC}^{\mathrm{final}}_{k,t} \right| \right) \\ R_{\mathrm{ol}} = \sum_{t\in\mathcal{T}} \sum_{k\in\mathcal{K}} 10 \cdot \frac{\left(P^{\mathrm{ch}}_{k,t} - P^{\mathrm{dis}}_{k,t}\right) - \overline{P}^{\mathrm{cs}}}{\overline{P}^{\mathrm{cs}}} \quad \text{if} > 0 \end{cases} \tag{35}$$

The PPO agent interacts with the environment and learns to optimize its policy by observing the rewards generated at each step. One epoch corresponds to a complete pass through the training data.

F. Hyperparameter Tuning and Simulation Settings

The hyperparameters of the RL model are set based on the values provided in [40], with the exception of the learning rate, which was set to 1×10^{-4}. During the pretraining phase, the model was trained over 10,000 epochs, while in the subsequent PPO training phase, the model was trained for 500,000 timesteps. The entire simulation model is based on Python 3.11.9. All simulations were conducted on a laptop featuring an Intel i7-11800H processor running at 2.30 GHz, with 32 GB of RAM.

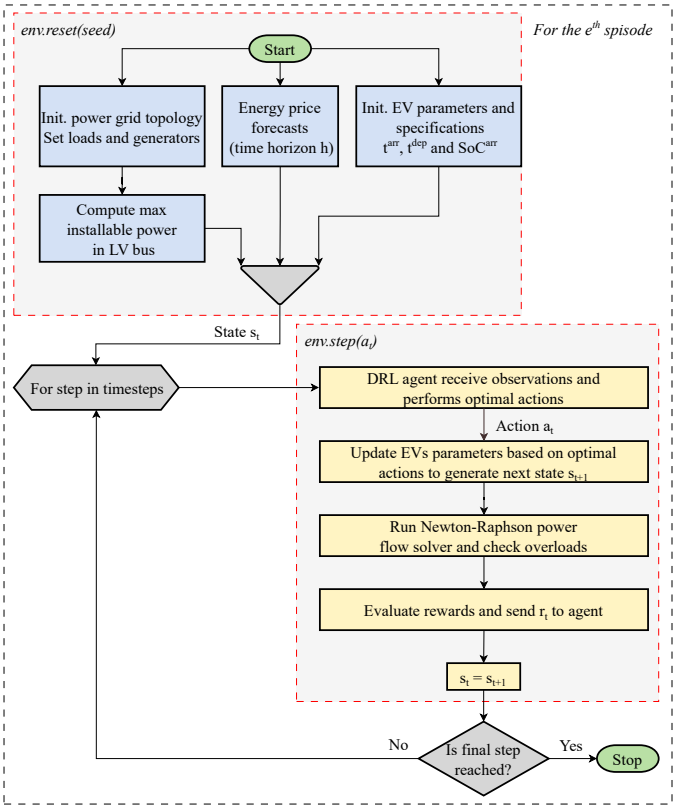


Fig. 5. Flowchart of the interaction of the environment with the agent.

V. Study Cases

In this section, the performances of the developed model are assessed by analyzing three different case studies. First, the general scenario is presented, where 500 distinct trajectories are analyzed to determine whether the RL algorithm achieves results comparable to those obtained using traditional optimization. Then, the analysis focuses on two specific case studies where the algorithm is put under stress: 1) a random-case scenario with a higher number of EVs during the same day, and 2) a worst-case scenario where all EVs arrive at the CSs when the grid is almost overloaded. In the latter two cases, the charging and discharging profiles of the EVs, as well as their impact on grid overload, are thoroughly examined. The charging and discharging actions obtained with the RL algorithm are compared to the optimal actions derived through an optimization algorithm using the Gurobi solver [41], under the assumption of full knowledge of external variables. Given this full-knowledge approach, an open-loop optimization is enough for the comparison, eliminating the need to solve the problem using a closed-loop strategy such as MPC. Both analyses are performed using 30-minute time steps and a 24-hour time horizon.

A. Case study 1: General-Case Scenario

In this subsection, we analyze the results of the general scenario in which the PPO algorithm, was tested on 500 random V2G service provision configurations. In this scenario, generic t^{arr} and t^{dep} as well as different power values were considered for each of the 500 trajectories. The only constant factor was the SOC^{target}, set at 80% of the nominal capacity for all cases. The performance of the PPO algorithm is then compared to a Traditional Optimization Algorithm (TOA) to assess how well the RL-based approach can handle these variable conditions.

The performance of the PPO algorithm is evaluated using several metrics, as shown in Table II. The results demonstrate that the PPO algorithm performs, in terms of accuracy, similarly to TOA across all key metrics. Specifically, user satisfaction, defined in (36), is almost identical for both methods, with PPO achieving a 97.3% satisfaction rate at CS 1 and 96.4% at CSs 2 and 3, compared to 100% for TOA.

$$\epsilon^{user} = \frac{1}{N} \cdot \sum_{n \in \mathcal{N}} \frac{SoC_n}{SoC_n^{target}} \tag{36}$$

The number of overloads and the overloading values are also nearly identical between the two methods, with only slight differences observed in the results. This is mainly because, considering the constraints into the loss function the problem is soft-constrained. All the constraints for the TOA are hard constraints, except for the station maximum power constraint described in (29), which has been set as a slack constraint. This is because it may be necessary to overload the grid temporarily to ensure that a vehicle can depart fully charged. Additionally, in cases where the grid is already overloaded, the contribution of the CSs might be insufficient to alleviate the situation. Consequently, while adhering to the station maximum power constraint is desirable, it is treated flexibly to avoid rendering the problem infeasible in scenarios where strict compliance is not possible. In terms of costs, PPO yields slightly higher values, with an average cost of approximately 50.718€ at station 1, as opposed to 44.278€ with TOA. The energy consumption results are similarly close, with PPO reporting an average energy consumption ranging from 147.6kWh to 149.9kWh, compared to TOA values of 157.1kWh to 157.3kWh. These small discrepancies indicate that PPO is able to manage the system effectively while providing slightly worse rewards for the users. However, the difference in performance between PPO and TOA is minimal, suggesting that PPO is a competitive solution for managing the charging process. It should also be noted that the reference solver operates with full knowledge of the environment, and is therefore considered the exact solution to the problem. On the other hand, the RL algorithm only knows the SoC of the EVs for the entire duration of their stay at the CS, the predictions of day-ahead market prices and the residual time before departure. This assumption is acceptable in light of recent developments in communication protocols for EVs in the context of providing services to the grid [42].

Turning now to the computational effort, the training of the NN involved two main phases. The pretraining phase took approximately 50 minutes to initialize the model, followed by the fine-tuning phase using PPO, which required an additional 4 hours. While these training times may seem substantial, it is important to note that they are incurred only once and for all. After the model is trained, it can be reused for multiple

TABLE II
USER SATISFACTION AND ALGORITHM PERFORMANCE METRICS.

Metric	CS	PPO		TOA	
		Mean [μ]	SD [σ]	Mean [μ]	SD [σ]
Avg. user sat [%]	1	97.3	8.289	100	0
	2	96.4	8.367	100	0
	3	96.4	8.367	100	0
N° Ovl.	Grid	5.125	3.228	5.207	3.745
Overload val. [%]	Grid	105.775	3.444	104.998	3.682
Cost [€]	1	50.718	19.926	44.278	16.996
	2	53.153	19.736	50.545	19.105
	3	52.954	19.071	50.689	18.556
Energy [kWh]	1	149.973	28.827	157.140	19.292
	2	147.599	28.819	157.273	19.042
	3	147.667	28.803	157.146	19.268

simulations, making the model highly efficient for real-time applications where quick responses are needed.

For the generation of the 500 trajectories, the RL algorithm required 1h30, which is notably faster than the 3h45 needed by TOA to solve the same problem. These times align with those reported in other studies, where traditional solvers based on mathematical models are generally more computationally intensive.

Despite the longer initial training time with PPO, the one-time computational cost makes it an attractive option for scenarios that require fast decision-making. Once the model is trained, it can generate results almost instantaneously, allowing for quicker responses in real-time applications. This efficiency, combined with similar performance outcomes compared to TOA, positions PPO as a viable and efficient solution for optimization problems in dynamic environments.

B. Case Study 2: Mean-Case Scenario

This subsection examines a generic scenario where the grid, initially excluding the presence of CSs, is not in an overload condition. In this scenario, the algorithm's primary focus is on maximizing profits from providing the V2G service, as power constraints are not a limiting factor. An important and noteworthy finding is that the learning process was carried out considering only one vehicle per day. However, as demonstrated by this case study, this assumption does not limit the model's performance. By fully randomizing the arrival and departure times of the vehicles, the algorithm can optimize the V2G service based solely on the SoC at each moment and the remaining hours until departure, effectively solving the optimization problem. As for the costs, there is an increase of 1.87% by the RL compared to TOA. Regarding user satisfaction ϵ^{user} defined in (36), in 50% of the cases it was above 100% (indicating a SoC higher than the target), and in the remaining 50% of the cases it never went below 80% of the target (i.e. 64% SoC in that specific case). The figures clearly show that the evolution of the SoC closely aligns with the TOA, and that the energy trading actions are designed to exploit the price profile to minimize costs. These results are illustrated in Fig. 6 for the SoC profiles and Fig. 7 for the corresponding action profiles.

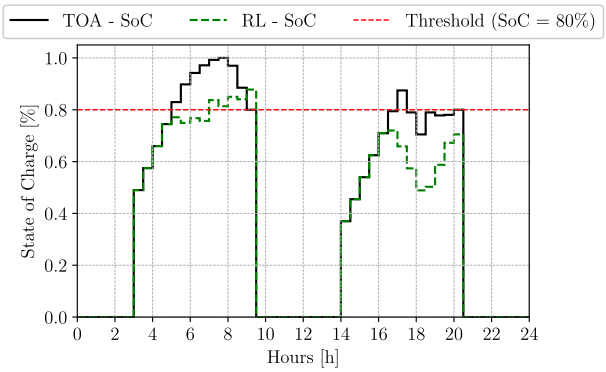


Fig. 6. SoC profiles in presence of more EVs during a single day.

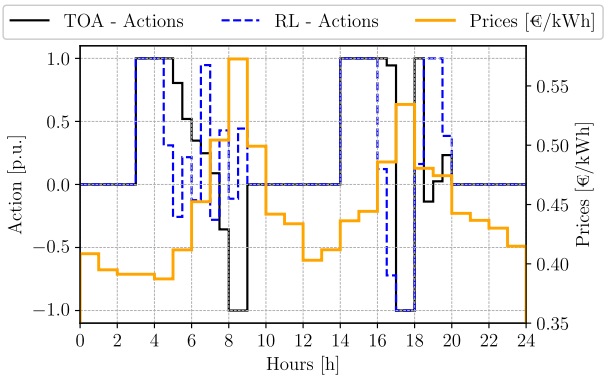


Fig. 7. Actions profiles in presence of more EVs during a single day.

C. Case Study 3: Worst-Case Scenario

This subsection considers the worst-case scenario, where the grid is already very close to an overload condition and all EVs arrive during peak hours. The presence of CSs in such a situation risks exacerbating the overload, pushing transformers dangerously close to 130% of their rated capacity. In this case, the controller must limit the cumulative power demand of the CSs. To demonstrate this, an ablation study was conducted in which only the charging and discharging of the vehicles was controlled based on electricity prices, without any consideration of the power limits required by the transformers. In Fig. 10, it can be seen that both the RL and the TOA model aim to regulate the power generation in order to avoid further overload. Specifically, it can be observed that for RL, the number of violations reached 4, with a maximum value of 108%, whereas with TOA, the number of violations was 7 but with smaller magnitudes, around 101%. Furthermore, as shown in Fig. 8, user satisfaction in this scenario reached a value close to 100%, indicating that the optimization process effectively manages the grid's stability while meeting the demands of the EVs. The corresponding actions are shown in Fig. 9.

VI. CONCLUSIONS AND FUTURE WORKS

In this work, we presented a novel RL-based aggregator model for managing V2G service provision within a distri-

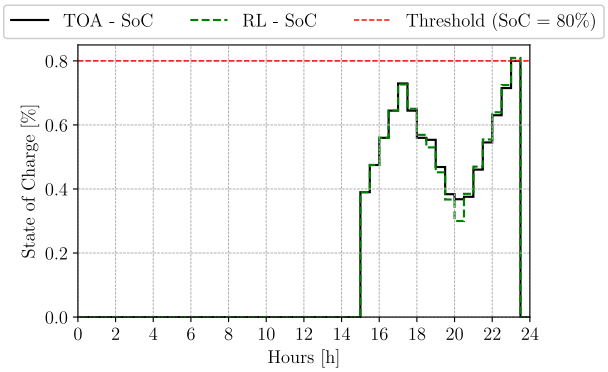


Fig. 8. SoC profiles in the worst case scenario.

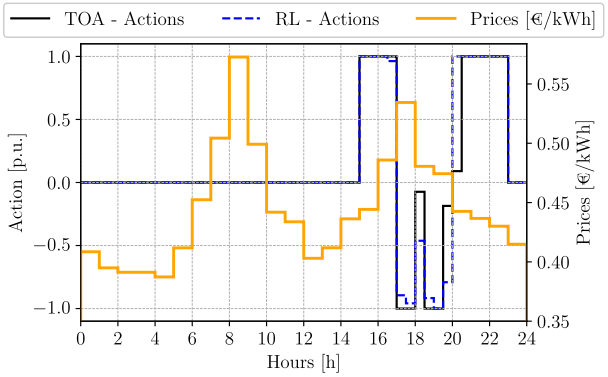


Fig. 9. Actions profiles in the worst case scenario.

bution network. The focus of the study was on optimizing the charging and discharging cycles of EVs to minimize costs, ensure user satisfaction and reduce network transformers overloads. The proposed model employed the PPO algorithm for RL to control the charging behavior of EVs, considering real-time network conditions and constraints. It also integrated a realistic equivalent circuit model for the EV battery and the model-inversion method to precisely calculate the battery operational limits.

The performance of the model was rigorously evaluated through a series of case studies. The results demonstrated that the RL-based approach, despite its reliance on limited information compared to traditional solvers, achieved performance comparable to TOA in terms of cost minimization, user satisfaction, and network overload management. In particular, the RL model operated efficiently in scenarios with dynamic conditions, such as varying EV arrival and departure times and grid congestion, while minimizing computational time. Furthermore, the results confirmed the model's ability to manage a variable number of EVs on the same day. Even when the network was near its capacity limits (worst-case scenario), the RL algorithm successfully regulated EV charging to prevent excessive transformer overloads and maintain grid stability.

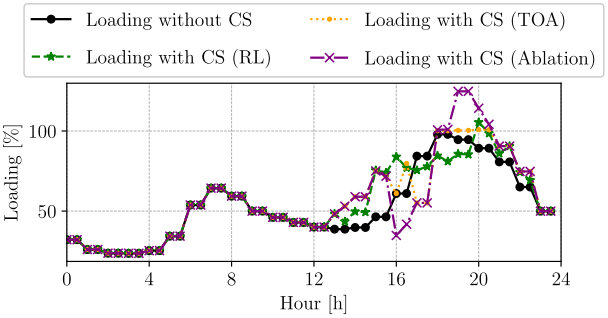


Fig. 10. Overloading profile throughout the day showing the performances of the different algorithms and considering the ablation study.

The RL-base approach proved to be highly competitive, offering a promising solution for real-time decision-making in dynamic grid environments. Additionally its ability to manage charging processes in a decentralized and adaptive manner makes it a strong candidate for large-scale EV integration into smart grids.

Future research could focus on developing hard-constrained models to ensure stricter adherence to grid limits, especially under high-load conditions. Additionally, exploring scenarios where the arrival and departure times of EVs are unknown would provide further challenges for the model, highlighting its ability to adapt to more uncertain and dynamic conditions. Another potential direction is the development of a multi-agent model for the coordinated management of multiple EVs. Another area for improvement in future work involves the challenges of training reinforcement learning (RL) algorithms in the real world, which arise from the high costs and risks associated with data collection. Innovative techniques can help bridge the gap between simulated experiences and real-world applications. By utilizing powerful simulation environments with domain randomization, targeted real-world data collection, and safe RL methods, researchers can overcome the limitations imposed by limited real-world data and enhance the transfer of policies from simulation to actual implementation. Key approaches include Domain Randomization and System Identification, which focus on adjusting simulation parameters and calibrating environments to develop robust policies. Furthermore, utilizing pre-collected historical interaction data enables the use of offline RL methods, reducing the need for continuous exploration in the real world. By integrating these strategies, future research can effectively address the challenges posed by limited real-world data and improve the transferability of trained policies.

REFERENCES

[1] International Energy Agency, "Global EV Outlook 2024," Paris, 2024, licence: CC BY 4.0. Accessed: 2024-01-30, 15:27. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2024

[2] R. A. Ibrahim, I. M. Gaber, and N. E. Zakzouk, "Analysis of multidimensional impacts of electric vehicles penetration in distribution networks," *Scientific Reports*, vol. 14, no. 1, p. 27854, November 2024.

[3] M. Topel and J. Grundius, "Load management strategies to increase electric vehicle penetration—case study on a local distribution network in stockholm," *Energies*, vol. 13, no. 18, 2020.

[4] A. Visakh and M. P. Selvan, "Analysis and mitigation of the impact of electric vehicle charging on service disruption of distribution transformers," *Sustainable Energy, Grids and Networks*, vol. 35, p. 101096, 2023.

[5] M. van den Berg, I. Lampropoulos, and T. AlSkaif, "Impact of electric vehicles charging demand on distribution transformers in an office area and determination of flexibility potential," *Sustainable Energy, Grids and Networks*, vol. 26, p. 100452, 2021.

[6] H. M. Al-Alwash, E. Borcoci, M.-C. Vochin, I. A. M. Balapuwaduge, and F. Y. Li, "Optimization schedule schemes for charging electric vehicles: Overview, challenges, and solutions," *IEEE Access*, vol. 12, pp. 32 801–32 818, 2024.

[7] Y. Wu, Z. Wang, Y. Huangfu, A. Ravey, D. Chrenko, and F. Gao, "Hierarchical operation of electric vehicle charging station in smart grid integration applications — an overview," *International Journal of Electrical Power Energy Systems*, vol. 139, p. 108005, 2022.

[8] O. Sadeghian, A. Oshnoei, B. Mohammadi-ivatloo, V. Vahidinasab, and A. Anvari-Moghaddam, "A comprehensive review on electric vehicles smart charging: Solutions, strategies, technologies, and challenges," *Journal of Energy Storage*, vol. 54, p. 105241, 2022.

[9] K. G. Firouzjah, "Profit-based electric vehicle charging scheduling: Comparison with different strategies and impact assessment on distribution networks," *International Journal of Electrical Power Energy Systems*, vol. 138, p. 107977, 2022.

[10] H. Yao, Y. Xiang, C. Gu, and J. Liu, "Optimal planning of distribution systems and charging stations considering pv-grid-ev transactions," *IEEE Transactions on Smart Grid*, vol. 16, no. 1, pp. 691–703, 2025.

[11] A. I. Aygun, M. S. Hasan, A. Joshi, and S. Kamalasadan, "A two-stage optimal electric vehicles charging methodology based on aggregators considering grid reliability and operational efficiency," *IEEE Transactions on Industry Applications*, vol. 61, no. 1, pp. 940–954, 2025.

[12] Z. Zhao, C. K. Lee, X. Yan, and H. Wang, "Reinforcement learning for electric vehicle charging scheduling: A systematic review," *Transportation Research Part E: Logistics and Transportation Review*, vol. 190, p. 103698, 2024.

[13] C. Zhang, Y. Liu, F. Wu, B. Tang, and W. Fan, "Effective charging planning based on deep reinforcement learning for electric vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 542–554, 2021.

[14] M. Dorokhova, Y. Martinson, C. Ballif, and N. Wyrsch, "Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation," *Applied Energy*, vol. 301, p. 117504, 2021.

[15] D. An, F. Cui, and X. Kang, "Optimal scheduling for charging and discharging of electric vehicles based on deep reinforcement learning," *Frontiers in Energy Research*, vol. 11, p. 1273820, 2023.

[16] H. M. Abdullah, A. Gastli, and L. Ben-Brahim, "Reinforcement learning based ev charging management systems–a review," *IEEE Access*, vol. 9, pp. 41 506–41 531, 2021.

[17] L. Yang, X. Geng, X. Guan, and L. Tong, "Ev charging scheduling under demand charge: A block model predictive control approach," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 2, pp. 2125–2138, 2024.

[18] C. Diaz-Londono, G. Fambri, P. Maffezzoni, and G. Gruosso, "Enhanced ev charging algorithm considering data-driven workplace chargers categorization with multiple vehicle types," *eTransportation*, vol. 20, p. 100326, 2024.

[19] C. Diaz-Londono, P. Maffezzoni, L. Daniel, and G. Gruosso, "Comparison and analysis of algorithms for coordinated ev charging to reduce power grid impact," *IEEE Open Journal of Vehicular Technology*, vol. 5, pp. 990–1003, 2024.

[20] S. H. Mansour, S. M. Azzam, H. M. Hasanien, M. Tostado-Véliz, A. Alkuhayli, and F. Jurado, "Deep reinforcement learning-based plug-in electric vehicle charging/discharging scheduling in a home energy management system," *Energy*, vol. 316, p. 134420, 2025.

[21] A. Das and D. Wu, "Optimal coordination of electric vehicles for grid services using deep reinforcement learning," in *2024 IEEE Power Energy Society General Meeting (PESGM)*, 2024, pp. 1–5.

[22] I. Azzouz and W. Fekih Hassen, "Optimization of electric vehicles charging scheduling based on deep reinforcement learning: A decentralized approach," *Energies*, vol. 16, no. 24, 2023.

[23] Y. Han, T. Li, and Q. Wang, "A dqn based approach for large-scale evs charging scheduling," *Complex Intelligent Systems*, vol. 10, no. 6, pp. 8319–8339, 2024.

[24] D. Liu, P. Zeng, S. Cui, and C. Song, "Deep reinforcement learning for charging scheduling of electric vehicles considering distribution network voltage stability," *Sensors*, vol. 23, no. 3, 2023.

[25] ACER, "Recommendation no 03/2023 of the european union agency for the cooperation of energy regulators of 19 december 2023," 2023, accessed: 14 March 2025. [Online]. Available: https://www.acer.europa.eu

[26] K. Strunz, "Developing benchmark models for studying the integration of distributed energy resources," in *2006 IEEE Power Engineering Society General Meeting*, 2006, pp. 2 pp.–.

[27] P. Community, "Cigre network example," 2024. [Online]. Available: https://pandapower.readthedocs.io/en/v2.1.0/networks/cigre.html

[28] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," 2017. [Online]. Available: https://arxiv.org/abs/1502.05477

[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[30] S. Orfanoudakis, C. Diaz-Londono, Y. E. Yılmaz, P. Palensky, and P. P. Vergara, "Ev2gym: A flexible v2g simulator for ev smart charging research and benchmarking," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2024.

[31] C. F. Lee, K. Bjurek, V. Hagman, Y. Li, and C. Zou, "Vehicle-to-grid optimization considering battery aging," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 6624–6629, 2023, 22nd IFAC World Congress.

[32] Y. Li, T. Wik, Y. Huang, and C. Zou, "Nonlinear model inversion-based output tracking control for battery fast charging," *IEEE Transactions on Control Systems Technology*, vol. 32, no. 1, pp. 225–240, 2024.

[33] C. D. Korkas, C. D. Tsaknakis, A. C. Kapoutsis, and E. Kosmatopoulos, "Distributed and multi-agent reinforcement learning framework for optimal electric vehicle charging scheduling," *Energies*, vol. 17, no. 15, 2024.

[34] L. Thurner, A. Scheidler, F. Schafer, J. H. Menke, J. Dollichon, F. Meier, S. Meinecke, and M. Braun, "pandapower - an open source python tool for convenient modeling, analysis and optimization of electric power systems," *IEEE Transactions on Power Systems*, 2018.

[35] F. Angizeh, A. Ghofrani, and M. A. Jafari, "Dataset on hourly load profiles for a set of 24 facilities from industrial, commercial, and residential end-use sectors," Available at:https://data.mendeley.com/datasets/rfnp2d3kjp/1, 2020, accessed: 2023-11-02.

[36] W. ur Rehman, I. A. Sajjad, T. N. Malik, L. Martirano, and M. Manganelli, "Economic analysis of net metering regulations for residential consumers in pakistan," in *2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC ICPS Europe)*, 2017, pp. 1–6.

[37] "Open power system data," https://open-power-system-data.org/, 2020, data obtained from ENTSO-E Transparency (Accessed: 11.12.2023).

[38] L. Lu, R. Pestourie, W. Yao, Z. Wang, F. Verdugo, and S. G. Johnson, "Physics-informed neural networks with hard constraints for inverse design," *SIAM Journal on Scientific Computing*, vol. 43, no. 6, pp. B1105–B1132, 2021. [Online]. Available: https://doi.org/10.1137/21M1397908

[39] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

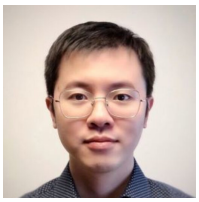
[40] A. Raffin, J. Kober, and F. Stulp, "Smooth exploration for robotic reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/2005.05719

[41] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: https://www.gurobi.com

[42] W. Terrance, K. Kouadio, and T. Youssef, "Understanding open charge point protocol," in *SoutheastCon 2023*, 2023, pp. 559–564.

Federico Rossi (Graduate Student Member, IEEE) was born in Udine, Italy, in March 1997. He received the B.S. degree in energy engineering from the Università degli Studi di Padova, in 2019, and the M.S. degree in electrical engineering from Politecnico di Milano, in 2022. He is currently pursuing the Ph.D. degree in electrical engineering at Politecnico di Milano. His research interests include the modeling of electrical power systems, digital twins, smart grids, and storage systems. Recently, his research has focused on reinforcement learning and optimization algorithms applied to power systems and electric mobility. He was a visiting PhD at Chalmers University of Technology, where he developed a machine learning based model for the efficient scheduling of electric vehicles charging and discharging in a vehicle-to-grid (V2G) context. Additionally, he was a visiting researcher at Polytechnique Montréal, where he developed a reinforcement learning algorithm for solving the optimal power flow problem.

Cesar Diaz-Londono (Senior Member, IEEE) received his B.Sc. and M.Sc. degrees in Electronics Engineering from Pontificia Universidad Javeriana, Colombia, in 2014 and 2016, respectively. He obtained a double doctoral degree: the Ph.D. in Engineering from Pontificia Universidad Javeriana and the Ph.D. in Electrical, Electronics, and Communications Engineering from Politecnico di Torino, Italy, in 2020. He worked as a postdoctoral researcher at Politecnico di Torino (2020–2021) and later as an Assistant Professor at Politecnico di Milano, Italy (2022–2024). In 2024, he held visiting research positions at Delft University of Technology, Netherlands, and Chalmers University of Technology, Sweden. Currently, he is a researcher at the Microgrid and Renewable Energy Research Center at Huanjiang Laboratory, Zhejiang University, China, and serves as the Young Professional Chair of the IEEE Transportation Electrification Council. His research interests include electric vehicle integration into power grids, demand response, smart grids, distributed energy resources, real-time electrical network simulation, and optimization.



Yang Li (Senior Member, IEEE) received the B.E. degree in electrical engineering from Wuhan University, Wuhan, China, in 2007, and the M.Sc. and Ph.D. degrees in power engineering from Nanyang Technological University (NTU), Singapore, in 2008 and 2015, respectively. He was a Research Fellow with the Energy Research Institute, NTU, and the School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD, Australia. He joined the School of Automation, Wuhan University of Technology, Wuhan, in 2019, as a Faculty Member. Since 2020, he has been a Researcher with the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden. His research interests include modeling and control of energy storage systems in power grid and transport sectors. Dr. Li serves as an Associate Editor for several journals such as IEEE Transactions on Industrial Electronics, IEEE Transactions on Transportation Electrification, and IEEE Transactions on Energy Conversion.

Changfu Zou (Senior Member, IEEE) received the Ph.D. degree in automation and control engineering from the University of Melbourne, Melbourne, VIC, Australia, in 2017. He was a Visiting Student Researcher at the University of California at Berkeley, Berkeley, CA, USA, from 2015 to 2016. Since early 2017, he has been with the Automatic Control Unit, Chalmers University of Technology, Gothenburg, Sweden, where he started as a Post-Doctoral Researcher and then became an Assistant Professor and is currently an Associate Professor. His research focuses on advanced modeling and automatic control of energy storage systems, particularly batteries. Dr. Zou has been awarded several prestigious grants from European Commission and Swedish national agencies and has hosted four researchers to achieve the Marie Skłodowska-Curie Fellows. He serves as an Associate Editor/Editorial Board Member for journals, such as IEEE Transactions on Vehicular Technology, IEEE Transactions on Transportation Electrification, and Cell Press journal iScience.

Giambattista Gruosso (Senior Member, IEEE) was born in 1973. He received the B.S, M.S, and Ph.D. degrees in electrical engineering from Politecnico di Torino, Italy, in 1999 and 2003, respectively. From 2002 to 2011, he was an Assistant Professor with the Department of Electronics and Informatics, Politecnico di Milano, where he has been an Associate Professor, since 2011. He is the author of more than 80 papers on journals and conferences on the topics. His research interests include electrical engineering, electronic engineering, industrial engineering, electrical vehicles transportation electrification, electrical power systems optimization, simulation of electrical systems, digital twins for smart mobility, factory and city, and how they can be obtained from data.