A Multiagent Deep Reinforcement Learning Based Approach for the Optimization of Transformer Life Using Coordinated Electric Vehicles

Li, Sichen; Hu, Weihao; Cao, Di; Zhang, Zhenyuan; Huang, Qi; Chen, Zhe; Blaabjerg, Frede

Published in:
I E E E Transactions on Industrial Informatics

DOI (link to publication from Publisher):
10.1109/TII.2021.3139650

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

A Multi-agent Deep Reinforcement Learning-Based Approach for the Optimization of Transformer Life Using Coordinated Electric Vehicles

Sichen Li, Weihao Hu, *Senior Member, IEEE*, Di Cao, *Student Member, IEEE*, Zhenyuan Zhang, *Senior Member, IEEE*, Qi Huang, *Fellow, IEEE*, Zhe Chen, *Fellow, IEEE*, and Frede Blaabjerg, *Fellow, IEEE*

***Abstract*— The uncertainties of charging behavior of EV owners have negative impact on the loss of life (LOL) of distribution transformer. This paper proposes a decentralized EV charging framework for optimization of the LOL of distribution transformer considering the dissatisfactions of EV owners. Specifically, long-short-term memory (LSTM) neural network is first utilized to capture the uncertainties caused by the load demand and electricity price. After that, each EV is modeled as an intelligent agent and a multi-agent deep reinforcement learning (MADRL) approach is applied to solve the coordinated charging problem based on the forecasting information by the LSTM network. All the agents are trained in a centralized manner to develop coordinated control strategies while inform decisions based on local information when finish the training process. The proposed approach can achieve coordinated scheduling of EVs based on local information, which helps preserve the privacy of EV owners, reduce the cost induced by the deployment of communication devices and avoid single-point failure. In addition, the parameter space noise and deep dense architecture in reinforcement learning are introduced to overcome premature convergence, training instability and inefficiency due to the large action space of multi-agent scenario. Comparative tests are carried out among several benchmark approaches utilizing real-world data to illustrate the effectiveness of the proposed approach.**

***Index Terms*—Multi-agent deep reinforcement learning; transformer life; EV charging scheduling**

I. Introduction

Owing to the gradual progress of related research, people realized that electric vehicle (EV) provided a feasible way to reduce problems related to air pollution and the depletion of conventional carbon energy sources. However, increasing penetration of EVs brings numerous technical challenges to the power grids, where the distribution network is the most vulnerable part to the adverse impact of unregulated EV charging [1]. In general, there are two main categories of negative impacts on distribution network caused by unregulated EVs charging: 1) system-level impacts and 2) equipment-level impacts [2]. The system-level impacts mean that the impacts of unregulated EVs charging on the power quality and power losses of distribution network system [3]. The equipment-level impacts represent the impacts of unregulated EVs charging on the asset of distribution network system (e.g., distribution line [4] and transformer [5]). This study mainly focuses on equipment-level, whose research necessary has been evaluated in studies [6], [7].

From the equipment-level perspective, ref. [8] points out that transformers are the distribution network assets most affected by the unregulated charging of EVs, as EV owners prefer to charge at home. Such negative impacts from unregulated charging EVs will lead to accelerated ageing of transformer and will have to replace it in advance to accommodate the additional power peaks required for EV loads. Aging is a serious problem faced by transformers and the main cause of the vast majority of transformer failures [9]. The aging of transformers is related to the aging of windings, bushings or on-load tap-changers, tanks, where the aging of windings means the influence of insulation material aging caused by winding hottest-spot temperature on insulation life of transformer, is the most concerned research direction [10]. The reduction of transformer insulation life usually refers to the "loss-of-life" (LOL), whose model is related to the aging acceleration factor and the normal insulation life [11]. Ref. [12] indicates that unregulated EVs charging, i.e., the lack of management of EVs charging, will cause the increase of transformer LOL. The reason is that unregulated EVs charging will cause transformers to bear excessive loads during peak charging periods, which will dramatically increase the winding hottest-spot temperature of transformers. Transformer insulation life is sensitive to the winding hottest-spot temperature and excessive temperature will cause accelerated insulation material aging which in turn reduces the insulating life of transformer. In contrast, transformer LOL will be dramatically reduced if the EVs charging are managed in coordination. In this context, the EVs coordinated charging management considering the effects of EVs on transformer LOL is a meaningful research topic. Generally, most of existing studies discusses two control architectures: centralized; and distributed or decentralized management approach [13].

Fully centralized approaches usually rely on a dedicated central controller which responses to collect the global data, perform calculation and optimization and determinate all control units' decision. For example, the work in [14] first collects the information of the whole EVs and then presented a fuzzy logic approach to schedule the EV charging to optimize the transformer LOL. The study in [8] investigates a centralized approach to manage the EVs charging, considering both EV owners' profits and the cost due to the transformer LOL. In [2], [15], the centralized-based control approach is proposed to maintain transformer LOL and service quality for EV owners. Several centralized charging

Manuscript received xxx; accepted xxx. Date of publication xxx; date of current version xxx. This work was supported by the National Key Research and Development Program of China under Grant 2018YFE0127600. Paper no. xxx. (*Corresponding author: Di Cao*.)

Sichen Li, Weihao Hu, Di Cao and Zhenyuan Zhang are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: sichenli@std.uestc.edu.cn; whu@uestc.edu.cn; caodi@std.uestc.edu.cn; zhangzhenyuan@uestc.edu.cn).

Qi Huang is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China. He is also with the College of Energy, Chengdu University of Technology, Chengdu, China. (e-mail: hwong@uestc.edu.cn)

Zhe Chen and Frede Blaabjerg are with the Department of Energy Technology, Aalborg University, DK-9220 Aalborg, Denmark (e-mail: zch@et.aau.dk; fbl@et.aau.dk).

schemes for EV in the workplace parking lot are implemented in [16] to optimize both economic benefits of parking lot operators and the life of transformers. Most recently, ref. [2] considers a co-optimization approach for both EV owners and transformer in distribution network, where the EV coordinated charging management formulated as optimization problem with the aim of minimizing energy losses and transformer's operation cost. From the above existing studies, we can conclude that the EV charging management system still has some deficiencies. Firstly, the works in [8] and [14] utilizes single EV charging management pattern to satisfy the whole EVs' requirements, where the requirements of different kinds of EVs are ignored. As the charging requirements for EVs vary depending on the demands of EV owners, applying the single charging management pattern for various EV owners may omit the difference between their demands. Given this, the EVs owners' optimization model should consider the variety of EV owners. Secondly, the real-time central controller mentioned in [2], [15], [16] need to update the information from EVs to manage their charging at each time step. Such a mechanism may result in the high bandwidth required for information exchange between central center and EVs, and even cause communication pressure if information is exchanged too frequently. Thirdly, if the problem to be solved has large-scale solution space, ref. [2] may be difficult to deploy online. More specifically, once the optimization is required, the approach presented in [2] must calculate the whole or part of the possible solutions and select the best one. This is a time-consuming process if the solution space is very large. Finally, the above-mentioned centralized-based approach rely on a powerful centralized controller, which needs to process a large amount of information at a single point simultaneously. In terms of EV coordinated charging management, this mechanism can easily lead to EV owners' privacy disclosure, the increased cost due to the deployment of communication devices and single-point failures. In order to mitigate the negative impact of traditional centralized approach on EV coordinated charging management to a certain extent, distributed and decentralized approach are effective alternatives.

One of the biggest differences between distributed and centralized approach is that the former no longer relies on central controller to optimize the objective function and thus avoids the single-point failures to a certain extent [17]. Recently, ref. [18] proposes a distributed-based multi-agent reinforcement learning to coordinate charging of EVs to ensure that the loads on transformers are below the limit values. To enhance the coordination and the performance of the approach, information are shared among multiple EVs. In addition to include the advantages of distributed approach, ref. [18] also easy to deploy online. However, demand differentiations among different EV owners are not taken into account, and there are still communication costs and risks of privacy disclosure.

Each controller in the decentralized approach does not need to construct communication channel with other controllers, which make decentralized approach not only keep the feature of the distributed approach, but also use only the local available information to execute the control. In total, the decentralized approach can bring three main advantages for the EVs charging management: 1) the privacy of each EV owner can be well protected; 2) the communication cost between each EV owners can be ignored; 3) compared with centralized approach, the stability is improved due to the single-point failures can be avoided. The coordinated charging relationship between multiple EVs leads to the coupling of charging decisions between them. In this case, any change in charging decisions of one EV will potentially influence the charging decisions of others, especially the influence will become more complex when the number of EVs increases. Although decentralized approaches for managing EV charging have many benefits mentioned above, the impact of charging decision coupling presents challenges to decentralized approach due to the coordination relationship is hard to construct through using only local information for decision making.

To overcome the challenge, this paper proposes a novel data-driven EV coordinated charging approach for the co-optimization of the transformer LOL and the different EV owners' demands based on multi-agent deep reinforcement learning (MADRL) approach with actor-critic framework. The proposed approach features the centralized training and decentralized execution to coordinate the EVs charging during the off-line training in centralized manner and deploy the charging management for each EV in on-line decentralized manner.

The main differences between the most of recently existing studies related to actor-critic MADRL algorithms and our proposed approach are as follows: the most of recently existing studies mainly focus on the alterations to the architecture of critic network to enhance the coordination among the whole agents, while ignoring the potential of actor network [19]-[23]. Unlike single-agent environments, the action space of multi-agent environments increases exponentially with the number of agents increasing, which easily causes premature convergence to local optimal solution. In order to avoid this problem, the parameter space noise (PSN) [24] is considered in the actor network to enhance the exploratory ability. Despite of the negative impacts associate with prematurely convergence can be solved by PSN, the steady performance and convergence speed of reward curves during training process may be influenced due to the introduction of noise. One possible solution is to introduce a technology that not only fits the PSN (without affecting the exploration of parametric noise), but also has strong non-linear representational power to accelerate the establishment of mapping relationship from input state to ideal output decision making. To ensure the training stability and efficiency, the actor network employs the deep dense architecture in reinforcement learning (D2RL) [25].

In this paper, each EV is modeled as an intelligent agent to coordinate with other EVs and make their own charging/discharging decisions during the charging period. To sum up, compared with the existing decision-making solutions, the main advantages of this paper are as follows:

(1) The coordinated charging of multiple EVs is formulated as a Markov game that is solved by the actor-critic MADRL approach features centralized training and decentralized execution. During the centralized training process, the critic network augmented with extra information help each agent explicitly model the decision process of other agents. This allows the actor network of each agent to achieve coordinated control even based on only local information at execution phase. In this way, the EVs coordinated charging management can be obtained in a completely decentralized manner.

(2) The critic network of the proposed approach utilizes the attention mechanism to process the whole EVs information and effectively guide the generation of coordinated strategies among actor network. In order to avoid converge prematurely to a local optimum due to the huge action space in multi-agent scenarios, each actor network utilizes PSN to construct effective exploration mechanism. Unlike traditional action space noise (ASN) added to action directly, the PSN can induce a consistent, complex, and state-dependent change in policy network over multiple time steps [24], [26]. In addition, the D2RL is deployed in actor network to improve the training instability and inefficiency caused by the

introduction of noise. PSN and D2RL are good combinations because PSN operates at the parameter level and D2RL builds a more complex and effective parametric connection relationship, which introduces greater potential for exploratory action than shallow dense architecture combined with PSN while eliminates the instability and inefficiency caused by noise during PSN exploration, resulting in achieving the better control performance.

(3) The proposed approach does not need to solve the complex EVs coordinated charging optimization problem in real-time. The decision-making functions of each agents can be constructed through offline training and be deployed online to select the control actions based on latest system state data.

The rest of this paper is organized as follows. Section II introduces the mathematical formulations of transformer LOL and EV owner's dissatisfaction model, and the optimization problem is reformulated as Markov games. In Section III the detailed descriptions of proposed approach are introduced. Section IV analyses the numerical simulation results. Finally, conclusions are discussed in Section V.

II. System Model And Problem Formulation

A. System description

Suppose there exists a community with a M kVA rated power distribution transformer and G EVs of different models. The distribution transformer is responsible for 1) the basic load (except EV load) of community residents and 2) the load of G EVs. In this paper, we assume the EV battery storage system allows transfer energy from or to the grid when the EVs are parked and plugged into the grid. In addition, it is assumed that the community is associated with only one utility company and as such the charging price of G EVs is the same at the same time. In this paper, each EV is considered as an agent, the goal of the G agents is to coordinate the charging of G EVs to reduce the transformer LOL and satisfy the EV owners' individual requirements under unknown load and electricity price of current moment. Normally, different EV owners have different commuting behaviors due to the different of individual habits, traffic [27], [28], etc. Thus, each episode starts at first EV arrival time and ends at last EV departure time, the length of an episode is $\max\{t_{dep}^{g} \mid g \in [1,G]\} - \min\{t_{arr}^{g} \mid g \in [1,G]\}$, where t_{arr}^{g} and t_{dep}^{g} represents the arrival time and departure time of g-th EV, respectively. Specially, the time interval Δt between two adjacent steps is assumed as one hour in this paper.

B. Transformer Loss of Life Model

The transformer lifetimes are related to winding hottest-spot temperature θ_{hs} according to insulation degradation model [29]. Before introducing the transformer LOL, the θ_{hs} should be known before. According to [29], θ_{hs} can be defined as:

$$\theta_{hs} = \theta_{amb} + \Delta\theta_{to} + \Delta\theta_{hs} \tag{1}$$

where the θ_{amb} is the average ambient temperature during the load cycle, $\Delta\theta_{to}$ represents top-oil rise over ambient temperature and $\Delta\theta_{hs}$ indicates the winding hottest-spot rise over top-oil temperature.

The $\Delta\theta_{to}$ and $\Delta\theta_{hs}$ can be calculated as [29]:

$$\Delta\theta_{to} = \left(\Delta\theta_{to,u} - \Delta\theta_{to,i}\right) \times \left(1 - e^{-du/\tau_{to}}\right) + \Delta\theta_{to,i} \tag{2}$$

$$\Delta\theta_{hs} = \left(\Delta\theta_{hs,u} - \Delta\theta_{hs,i}\right) \times \left(1 - e^{-du/\tau_{w}}\right) + \Delta\theta_{hs,i} \tag{3}$$

where du is the duration of load L. $\Delta\theta_{to,u}$ denotes the ultimate top-oil rise over ambient temperature for load L, $\Delta\theta_{to,i}$ is the initial top-oil rise over ambient temperature, $\Delta\theta_{hs,u}$ is the ultimate winding hottest-spot conductor rise over top-oil temperature for load L, $\Delta\theta_{hs,i}$ is the initial winding hottest-spot rise over top-oil temperature, τ_{to} and τ_{w} represent the oil time constant and winding time constant, respectively. The detailed explanations of the two constants can be found in [29].Therein, the $\Delta\theta_{to,u}$, $\Delta\theta_{to,i}$, $\Delta\theta_{hs,u}$ and $\Delta\theta_{hs,i}$ are given by the following equations [29]:

$$\Delta\theta_{to,u} = \Delta\theta_{to,r} \times \left[\frac{K_u^2 R + 1}{R + 1}\right]^n, \ \Delta\theta_{to,i} = \Delta\theta_{to,r} \times \left[\frac{K_i^2 R + 1}{R + 1}\right]^n \tag{4}$$

$$\Delta\theta_{hs,u} = \Delta\theta_{hs,r} \times K_u^{2m} \tag{5a}$$

$$\Delta\theta_{hs,i} = \Delta\theta_{hs,r} \times K_i^{2m} \tag{5b}$$

where $\Delta\theta_{to,r}$ is the top-oil rise over ambient temperature at rated load, $\Delta\theta_{hs,r}$ denotes the winding hottest spot conductor rise over top-oil temperature at rated load, R is the ratio of rated load loss to no-load loss, K_u is the ratio of ultimate load L to rated load, K_i is the ratio of initial load L to rated load, n is an empirically derived exponent used to calculate the variation of $\Delta\theta_{to}$ with changes in load, and m is an empirically derived exponent used to calculate the variation of $\Delta\theta_{hs}$ with changes in load.

In order to obtain the transformer LOL, the aging acceleration factor F_{AA} and equivalent aging factor F_{EQA} are necessary. The F_{AA} and F_{EQA} can be expressed as [11]:

$$F_{AA} = \exp\left(\frac{15000}{383} - \frac{15000}{\theta_{hs} + 273}\right), \ F_{EQA} = \frac{\sum_{k=1}^{N} F_{AA,k} \Delta t}{\sum_{k=1}^{N} \Delta t} \tag{6}$$

where N is the total number of time intervals, k denotes index, $F_{AA,k}$ is aging acceleration factor for the k-th time interval.

Finally, the transformer LOL is defined as:

$$LOL = \frac{F_{EQA} \times du}{\mathrm{NLL}} \tag{7}$$

where NLL means normal insulation life of transformer.

C. The Dissatisfactions of EV Owners Model

EV owners are selfish and usually charge EVs in the way of maximizing their own interests and as such it is impractical to ignore the dissatisfactions of EV owners in the coordinated charging problems [30]. From EV owner's perspectives, the dissatisfaction model is constructed by three objectives, including one objective with unity and two objectives with variety:

1) The cost of EV charging: Among EV owners, the cost due to EV charging is the objective with unity. The charging cost of g-th EV at time t can be defined as:

$$C_t^g = \mathcal{P}_t \cdot power_t^g \cdot \Delta t \tag{8a}$$

$$-\max(power^g) \le power_t^g \le \max(power^g) \tag{8b}$$

where $\mathcal{P}_t$ is the average electricity price during time t, $Power_t^g$ is the charging/discharging power of g-th EV at time t, and $-\max(power^g)$ and $\max(power^g)$ are the allowed maximum discharging and charging power of g-th EV, respectively. $power_t^g$ is positive for EV charging and negative for EV discharging. Utilizing the fluctuation of

electricity price, charging at high electricity price and discharging at low electricity price can make EV owners get economic benefit.

2) Range anxiety: Range anxiety is a measure of EV owner's concern that the EV does not have enough energy to reach its destination [28], [31]. One important point is that EV charging management should incorporate the variety of human mentalities. These individuals tend to have different range anxieties due to the various individual demands. In order to include richer and more complete scenarios, it is of important to differentiate between EV owners' range anxieties by different unique mathematical descriptions. In view of this, three mathematical descriptions of range anxieties $RA_{i=1,2,3}$ are considered in this paper, and the range anxiety of g-th EV owner is defined as follows,

$$RA_1^g = \frac{\left|E_{\max}^g - E_{t_{dep}}^g\right|}{E_{\max}^g}, RA_2^g = \left(\frac{E_{\max}^g - E_{t_{dep}}^g}{E_{\max}^g}\right)^2, RA_3^g = \frac{\ell(E_{t_{dep}}^g) + \left|\ell(E_{\max}^g)\right|}{\ell(0) + \left|\ell(E_{\max}^g)\right|} \quad (9)$$

where $E_{\max}^g$ denotes the max capacity of g-th EV, $E_{t_{dep}}^g$ denotes the energy of g-th EV at t_{dep}^g, $E_{\max}^g - E_{t_{dep}}^g$ represents the part of uncharged battery energy, and $\ell(E_{t_{dep}}^g) = \mathrm{In}\left(1\Big/\left(1.01 - \left(\left|E_{\max}^g - E_{t_{dep}}^g\right|\Big/E_{\max}^g\right)^2\right)\right)$. To visualize their differences, the $RA_{i=1,2,3}$ varies with state-of-charge (SOC) as shown in Fig. 1 where $SOC_{t_{dep}} = E_{t_{dep}} / E_{\max}$. In the Fig. 1, we can give a physical meaning to the absolute value of curve slope, that is, the range anxiety relief rate (RARR). With the same small $SOC_{t_{dep}}$ increment added, the larger the RARR means the greater the reduction in range anxiety. RARR is the indication of the variety of range anxiety among EV owners. In Fig. 1, there is a linear correlation between RA_1 and $SOC_{t_{dep}}$. Here, the constant slope indicates that the RA_1 maintain a constant RARR with the increase of $SOC_{t_{dep}}$. The constant RARR means that the EV owners with type RA_1 range anxiety maintains the uniform decrease relationships with the increase of $SOC_{t_{dep}}$. Where RA_2 differs RA_1 is that the former has the dynamic RARR. RA_2 has the biggest RARR at $SOC_{t_{dep}} = 0$ and RARR generally decreases as $SOC_{t_{dep}}$ increases and ultimately obtains the minimum value at $SOC_{t_{dep}} = 1$. The decrease of RARR with the increase of $SOC_{t_{dep}}$ indicates that when the same increment of $SOC_{t_{dep}}$ acts on the range of small $SOC_{t_{dep}}$, it can more significantly reduce range anxiety than when it acts on the range of large $SOC_{t_{dep}}$. Be similar to RA_2, RA_3 also has the dynamic RARR but the EV owners with RA_3 more easily satisfy the adequacy of battery energy than those with RA_2 due to RA_3 curve has the larger RARR within the range of small $SOC_{t_{dep}}$. This difference indicates that EV owners with RA_3 have shorter daily driving distances than EV owners with RA_2 and therefore require less energy, so the same $SOC_{t_{dep}}$ increment can have a greater effect on reducing range anxiety at the range of small $SOC_{t_{dep}}$ for the EV owners with RA_3 than for EV owners with RA_2. In addition, due to the smaller energy required, a large $SOC_{t_{dep}}$ is of less significance to EV owners with RA_3, and thus RARR of EV owners with RA_3 is smaller than EV owners with RA_2 in the case of large $SOC_{t_{dep}}$.

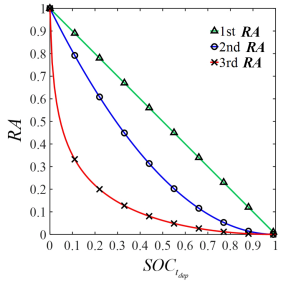


Fig. 1. The dynamic change of the $RA_{i=1,2,3}$ with $SOC_{t_{dep}}$.

3) The cost of battery degradation: The chemical cell degradation depends on the construction (e.g. battery and fuel cell) and the chemistry (e.g. lithium-ion battery, nickel-metal hydride battery, lithium iron phosphate battery and so on). This paper focus on the degradation of lithium-ion battery because the high density and high efficiency which the lithium-ion batteries have and as such widely used in EVs. Assuming the lithium-ion batteries are only sensitive to the total number of cycles as recommended in ref [32]. At time t, the cost of battery degradation of g-th EV is estimated as:

$$BD_t^g = \mathcal{C}^E \left|\frac{\Upsilon}{100}\right| \frac{power_t^g}{E_{\max}^g} \Delta t \quad (10)$$

where $\mathcal{C}^E$ is the total battery cost and it is different among EV owners, Υ denotes the slope of the linear approximation of the battery life as a function of the cycles.

D. Problem Reformulation

Among the total G EVs, the objective function of the g-th EV at time t can be defined as:

$$\zeta_t^g = \begin{cases} \mathrm{W}_{\mathrm{ra}} \cdot RA_t^g + C_t^g + BD_t^g + \dfrac{\mathrm{W}_{\mathrm{LOL}} \cdot \left(LOL_t^{total} - LOL_t^{basic}\right)}{G}, t = t_{dep} - 1 \\ C_t^g + BD_t^g + \dfrac{\mathrm{W}_{\mathrm{LOL}} \cdot \left(LOL_t^{total} - LOL_t^{basic}\right)}{G}, t \neq t_{dep} - 1 \end{cases} \quad (11)$$

$$s.t. \ -\max(power^g) \le power_t^g \le \max(power^g)$$

where W_{ra} is weighting factor measured in \$ to map the RA_t^g into money [31]. LOL_t^{basic} represents the LOL under the basic load $load_t^{basic}$ during time t, LOL_t^{total} denotes the LOL under the total load $load_t^{total}$ in the same time period and $\mathrm{W}_{\mathrm{LOL}}$ is the economic value of transformer, which is used to map the LOL into the economic benefit decrease of EV owners due to the LOL. Therein, the total load $load_t^{total}$ is defined as:

$$load_t^{total} = load_t^{basic} + \sum_{i=1}^{G} power_t^i \Delta t \quad (12)$$

where the $\sum_{i=1}^{G} power_t^i \Delta t$ denotes the load of all G EVs $load_t^{EV}$. Therefore, $LOL_t^{total} - LOL_t^{basic}$ indicates the LOL under the load of all G EVs.

Then, the charging of the total G EVs in an episode can be optimized by:

$$\mathcal{J} = \min \sum_{g=1}^{G} \sum_{t}^{K^g} \zeta_t^g \quad (13)$$

In order to solve $\mathcal{J}$, the EV coordinated charging problem can be considered as a multi-agent setting, in which each EV is considered as an agent. Then, the multi-agent setting of coordinated charging problem are reformulated as a Markov games [33]. Finally, a MADRL-based charging management is designed in this work to solve it.

In this paper, the electricity price $\mathcal{P}_t$ and basic load (except EV load) $load_t^{basic}$ are assumed as the unknown value due to the randomness of the electricity market. Thus, the $\mathcal{P}_t$ forecast value $\widetilde{\mathcal{P}}_t$ and the $load_t^{basic}$ forecast value $\widetilde{load}_t^{basic}$ are introduced in this model. Under this assumption, the forecast value of Eq. (5a) should be defined as:

$$\Delta\tilde{\theta}_{hs,u} = \Delta\theta_{hs,r} \times \widetilde{K}_u^{2m} \tag{14}$$

where K_u^{2m} is influenced by $load_t^{basic}$ [11]. The $\Delta\tilde{\theta}_{hs,u}$ indicates the forecast of $\Delta\theta_{hs,u}$ due to the $\widetilde{K}_u^{2m}$ is influenced by $\widetilde{load}_t^{basic}$. Basically on this assumption, the detail designing of this Markov games with G EVs are as follows:

1) State: At time t, the state of g-th EV is defined as:

$$s_t^g = (\tau^g,\ \theta_{hs,t},\ load_{t-1}^{EV},\ \widetilde{\mathcal{P}}_t,\ t,\ SOC_t^g,\ t_{dep}^g,\ SOC_{t_{dep}}^g) \tag{15}$$

where τ denotes the EV type and τ^g is the g-th EV type, $\theta_{hs,t}$ is the winding hottest-spot temperature at time t and it is influenced by the forecast of basic load $\widetilde{load}_t^{basic}$ due to the Eq. (14), $\widetilde{\mathcal{P}}_t$ indicates the forecast of $\mathcal{P}_t$, SOC_t^g is the state-of-charge (SOC) of g-th EV at time t, and $SOC_{t_{dep}}^g$ denotes the SOC at departure time.

2) Action: Given the state s_t^g, the action denotes the charging/discharging power of g-th EV, i.e., $a_t^g = power_t^g$.

3) Reward Function: We called the $\zeta_t^i,\ i \in [1, G]$ as individual reward of each agent, and $\sum_{i=1}^{G}\zeta_t^i$ is the reward of sum of each agent's individual reward. For an EV g, the reward of g-th EV is defined as: $r_t^g = -\sum_{i=1}^{G}\zeta_t^i$. It means that during the training process of g-th EV, the g-th EV should not only consider its own reward, but also consider the overall benefits.

4) Transition Function: For g-th EV, the state transition can be defined as: $s_{t+1}^g = \mathcal{T}\left(s_t^g, a_t^g, \vartheta_t\right)$. As shown by the transition function $\mathcal{T}$, there are two factors action a_t^g and randomness ϑ_t determine the state s_t^g to next state s_{t+1}^g in this model. It means that the state transition is not only controlled by the a_t^g but also influenced by the ϑ_t. In order to describe the state transition clearly, three parts can be divided: ***(a)*** Only controlled by a_t^g : the deterministic part is only controlled by a_t^g, i.e., battery model: $SOC_{t+1}^g = SOC_t^g + a_t^g \Delta t / E_{\max}^g$ and sum of the G EVs loads up to time t+1 : $\mathcal{L}_{t+1} = \mathcal{L}_t + \sum_{i=1}^{G} a_t^i \Delta t$. ***(b)*** Only influenced by ϑ_t : ϑ_t is used to represent the stochastic factor in the system. In this work, ϑ_t are the forecasted error of electricity price and load, and EV owner's commuting behavior. ***(c)*** Determining jointly by a_t^g and ϑ_t : according to ref [11], the hottest-spot temperature θ_{hs} at time t can be calculated by:

$$\theta_{hs,t} = \mathcal{F}\left(load_{t-q}^{total}, ..., load_{t-1}^{total}, \widetilde{load}_t^{basic}\right) \tag{16a}$$

and its value at time $t+1$ can be calculated by:

$$\theta_{hs,t+1} = \mathcal{F}\left(load_{t-q}^{total}, ..., load_{t-1}^{total}, \widetilde{load}_t^{total}\right) \tag{16b}$$

where $\mathcal{F}$ denotes the calculating process from Eq. (1) to Eq. (4) and Eq. (14).

III. Proposed Approach

As shown in Fig. 2, the proposed approach is comprised of two parts, 1) LSTM-based NN is trained for forecasting the future electricity price and EV owners' basic load, 2) a MADRL-based approach is developed for making real-time coordinated charging/discharging decisions after receiving the information from LSTM-based NN.

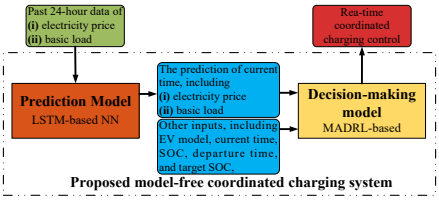


Fig. 2. The architecture of the proposed approach.

A. LSTM-based NN for Price and Load Forecasting

Algorithm 1: Training process of LSTM-based approach

Inputs: The past 24-hour electricity price $\mathcal{P}_{t-24}$, ..., $\mathcal{P}_{t-1}$ and basic load $load_{t-24}^{basic}$, ..., $load_{t-1}^{basic}$.

Outputs: The forecasted electricity price $\widetilde{\mathcal{P}}_t$ and basic load $\widetilde{load}_t^{basic}$ of current time.

1: **for** episode =1:N **do**

2: ▷ $\widetilde{\mathcal{P}}_t = \mathrm{LSTM}_1(\mathcal{P}_{t-24}, ..., \mathcal{P}_{t-1})$

$\widetilde{load}_t^{basic} = \mathrm{LSTM}_2(\widetilde{load}_{t-24}^{basic}, ..., \widetilde{load}_{t-1}^{basic})$

3: ▷ $Z^{\widetilde{\mathcal{P}}_t} = \left(\widetilde{\mathcal{P}}_t - \mathcal{P}_t\right)^2$

$Z^{\widetilde{load}_t^{basic}} = \left(load_t^{basic} - \widetilde{load}_t^{basic}\right)^2$

4: ▷ Updating the parameters of LSTM_1 and LSTM_2 to minimize the $Z^{\widetilde{\mathcal{P}}_t}$ and $Z^{\widetilde{load}_t^{basic}}$, respectively

5: **end for**

Recently, due to LSTM-based NN is comparatively easy to implement and shows good performance, the LSTM-based NN has been attracted a lot of attention in forecasting electricity price and load [34], [35]. The LSTM-based NN has two features on time sequential forecasting, one is cell state, which has a recurrent self-connected edge with a constant weight of 1 to overcome gradient disappearance and gradient explosion [36]; the other is gating mechanism that can selectively control the data flow through the gate, which can save the important and forget the useless feature of the sequential information [37]. In order to deal with the uncertainties of the unknown price and load, the LSTM-based NN is used to dynamically forecast the electricity price and basic load. Specifically, at each time step t, the inputs of the LSTM-based NN are past 24-hour (t-24, …, t-1) electricity price and EV owners' basic load, and its output are the forecasted current one hour (t) electricity price and EV owners' basic load, respectively. After that, the forecasted information will be fed into MADRL-based proposed approach to coordinate different EVs charging/discharging.

The training process of LSTM-based is summarized in Algorithm 1. In Algorithm1, the LSTM trains in traditional supervised manner. Obtain the predicted price and load at step 2, calculate the loss with the label data at step 3, and update the parameters of the two LSTM neural networks at step 4.

B. MADRL-based Approach for Decision-Making

Inspired by [38], the critic is utilized the attention module to enhance the coordination of multiple EVs, and applied PSN [24] and D2RL [25] in actor to establish an effective exploration mechanism.

The proposed approach is consisted by two neural

network based parts, one is actor part which responses to make decision; the other is critic part which responses to guide the actor part to approximate the optimal policy. The critic neural network consists of Q network Q^i and Q target network $Q^i_{\text{target}}, i \in [1, G]$. In the similar way, the actor neural network can be divided into policy network π^i and policy target network $\pi^i_{\text{target}}, i \in [1, G]$.

Critic network: the critic network utilized the attention mechanism to manage the whole EVs information, effectively guiding the generation of coordinated strategies among actor network.

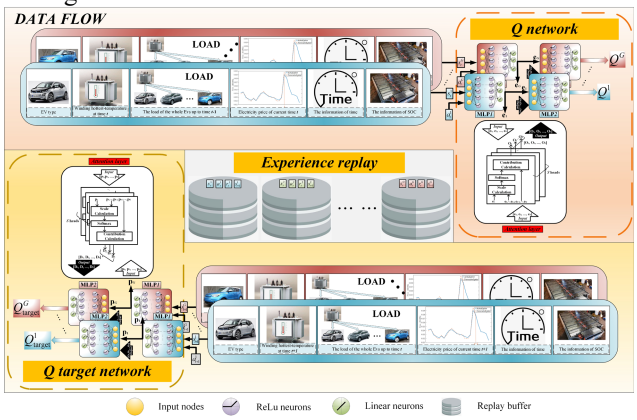


Fig. 3. Data flow of critic part.

The data flow of critic part is shown in Fig. 3, which is divided into two parts: Q network and Q target network. As shown in Q network part of Fig. 3, for each agent g, the concatenation of s_t^g and a_t^g input to MLP*1* and output $\mathbf{e}_g$. When the calculations of $\mathbf{e}_i, i \in [1, G]$ are finished, the concatenation $[\mathbf{e}_1, ..., \mathbf{e}_G]$ are input to the attention layer. The idea behind the attention layer is to enable the agent to focus on important information and suppress the impact of irrelevant details on the decision. Details of the attention layer are as follows: the mechanism of multiple attention heads [39], specifically, attention layer with S heads are used in the proposed approach. In each head, there are six optimizable parameters $\mathbf{w}_Q^i, \mathbf{b}_Q^i, \mathbf{w}_K^i, \mathbf{b}_K^i, \mathbf{w}_V^i, \mathbf{b}_V^i, i \in [1, S]$ and as such the total parameters of attention layer of proposed approach are $\underbrace{\{\mathbf{w}_Q^1, \mathbf{b}_Q^1, \mathbf{w}_K^1, ..., \mathbf{b}_K^S, \mathbf{w}_V^S, \mathbf{b}_V^S\}}_{6\times S}$. In the process of calculating the action-value Q^j of a specific agent j, each head i has the same input $[\mathbf{e}_1, ..., \mathbf{e}_G]$ to calculate with its own $\mathbf{w}_Q^i, \mathbf{b}_Q^i, \mathbf{w}_K^i, \mathbf{b}_K^i, \mathbf{w}_V^i, \mathbf{b}_V^i$ to get the final output O_i. After the calculation of S heads are completed, the calculation results are concatenated as the attention feature of an agent $[\mathrm{O}_1, \mathrm{O}_2, ..., \mathrm{O}_S]$. Note that, although the six optimizable parameters between each head are independent of each other, all different heads have the same input. This mechanism makes the head similar to the concept of convolution kernel in convolution neural network [40]. Another point that should also be noted is that the six optimizable parameters in each head i are shared across all G agents, which encourages a common embedding space [38]. Take the g-th agent as an example, the output of the specific head O_i with parameters $\mathbf{w}_Q^i, \mathbf{b}_Q^i, \mathbf{w}_K^i, \mathbf{b}_K^i, \mathbf{w}_V^i, \mathbf{b}_V^i$ can be calculated by:

(a) Scaled Calculation: For the g-th agent, $\mathbf{e}_g$ is calculated with $\mathbf{w}_Q^i$ and $\mathbf{b}_Q^i$, $[\mathbf{e}_1, ..., \mathbf{e}_{g-1}, \mathbf{e}_{g+1}, ..., \mathbf{e}_G]$ is calculated with $\mathbf{w}_K^i$ and $\mathbf{b}_K^i$ in this part:

$$\mathcal{I} = \frac{f\left(\mathbf{e}_g \mathbf{w}_Q^i + \mathbf{b}_Q^i, \left[\mathbf{e}_1 \mathbf{w}_K^i + \mathbf{b}_K^i, ..., \mathbf{e}_G \mathbf{w}_K^i + \mathbf{b}_K^i\right]\right)}{\sqrt{d_k}} = \left[\partial_1, ..., \partial_{g-1}, \partial_{g+1}, ..., \partial_G\right] \tag{17}$$

where $f(a, [b, ..., z]) = [a \cdot b, ..., a \cdot z]$, the d_k denotes the dimensions of $\mathbf{w}_Q^i$ and $\mathbf{w}_K^i$, and $\sqrt{d_k}$ is used to scaled calculation [39].

(b) Softmax: The attention of the g-th agent to other agents is calculated in this part. The attention weights of the g-th agent to other agents can be calculated as:

$$\omega_j = \frac{\exp(\partial_j)}{\sum_{j=1}^{G} \exp(\partial_j)}, j \in [1, ..., g-1, g+1, ..., G] \tag{18}$$

(c) Contribution Calculation: For the g-th agent, the contribution from other agents is a weight sum:

$$\mathrm{O}_i = \sum_{j=1}^{G} \omega_j \left(\mathbf{e}_j \mathbf{w}_V^i + \mathbf{b}_V^i\right), \quad j \in [1, ..., g-1, g+1, ..., G], \ i \in [1, S] \tag{19}$$

After the calculation of S heads are completed, the calculation results are concatenated as the attention feature of an agent $[\mathrm{O}_1, \mathrm{O}_2, ..., \mathrm{O}_S]$ and input to the MLP*2*.

Finally, the Q^g can be obtained by the output of MLP*2*, which is the mapping of the concatenation of $\mathbf{e}_g$ and $[\mathrm{O}_1, \mathrm{O}_2, ..., \mathrm{O}_S]$. The calculation process of Q are same for each agent.

Due to the optimizable parameters of attention layer are shared across all G agents, the optimizable parameters of each $Q^i, i \in [1, G]$ are updated together to minimize a joint regression loss function:

$$\text{Loss} = \sum_{i=1}^{G} \mathbb{E}_{\left(s_k^{AG}, a_k^{AG}, r_k^i, s_{k+1}^{AG}\right) \sim \mathcal{D}} \left[\left(Q^i\left(s_k^{AG}, a_k^{AG}\right) - y^i\right)^2\right] \tag{20a}$$

$$y^i = r_k^i + \gamma \mathbb{E}_{a_{k+1}^{AG} \sim \pi_{\text{target}}^{AG}\left(s_{k+1}^{AG}\right)} \begin{bmatrix} Q_{\text{target}}^i\left(s_{k+1}^{AG}, a_{k+1}^{AG}\right) \\ -\alpha \log\left(\pi_{\text{target}}^i\left(a_{k+1}^i \mid s_{k+1}^i\right)\right) \end{bmatrix} \tag{20b}$$

where $\mathcal{D}$ represents all G agents' experience replay; s_k^{AG} and a_k^{AG} denote states and actions for the all G agents at time step k, respectively, where $s_k^{AG} = \{s_k^1, ..., s_k^G\}$ and $a_k^{AG} = \{a_k^1, ..., a_k^G\}$; r_k^i is the reward for agent i at time step k; s_{k+1}^{AG} and a_{k+1}^{AG} denote states and actions for the all G agents at time step k+1, respectively; γ is the reward discount factor which is used to balance the immediate and future reward. π_{target}^{AG} is the policy target network for the all G agents, where $\pi_{\text{target}}^{AG} = \{\pi_{\text{target}}^1, ..., \pi_{\text{target}}^G\}$; $-\log\left(\pi_{\text{target}}^i\left(a_{k+1}^i \mid s_{k+1}^i\right)\right)$ indicates the entropy regularization, which is used to encourage exploration [41]; α is the temperature parameters, which is used to balance the exploration and exploitation during the training process.

The optimizable parameters of Q target network are updated by soft update mechanism [41].***Actor network:*** In the multi-agent environments, any increase in the number of agents will expand the action space in an exponential manner. To prevent premature convergence in the case of large-scale action space, exploration mechanism is very important for RL during the optimization. Generally, traditional deep reinforcement learning (DRL) utilized ASN to avoid the local optimum. For example, considering the continuous action space influenced by the Gaussian noise

$\mathcal{N}$ case, the action can be represented as $a_t = \pi(s_t) + \mathcal{N}$. In contrast with ASN, the PSN deploys noise to the parameter space of the policy network rather than the action space. If it is influenced by the PSN, the action can be denoted as $a_t = \tilde{\pi}(s_t)$, where the π is affected by the PSN to get $\tilde{\pi}$. Comparing ASN and PSN, the former is completely independent of the s_t since even the fixed s_t it is, the a_t is not the same due to the ASN. In other words, the ASN belongs to the state-independent mechanism. In cases where ASN has great impact on actions, even though training can establish a mapping relationship between states and actions, such a mechanism is likely to weaken the dependency between states and actions, resulting in poor performance in complex environments [42]. However, even with these potential drawbacks, ASN is still the most widely used and popular choice for DRL today. The primary reason for this is the lack of easy to deploy, computational cheap exploration approaches to avoid the above-mentioned problems while inducing policy network to achieve good results. The PSN aims to fill the need. The PSN acts on the parameters of policy network at each episode and kept itself fixed until the next episode comes. This ensures consistency in actions, and directly introduces a dependence between the state and the exploratory action taken [24], [26]. In order to describe the PSN clearly, it is assumed that there is a linear layer of a neural network with shape p inputs and shape q outputs, represented by:

$$y = \omega x + b \tag{21}$$

where $y \in \mathbb{R}^q$ is the output, $x \in \mathbb{R}^p$ represents the input, $\omega \in \mathbb{R}^{q \times p}$ denotes the parameters, and $b \in \mathbb{R}^q$ represents the bias. The corresponding layer with PSN can be defined as:

$$y = \left(\omega + \mathcal{N}^{\omega}\right) x + \left(b + \mathcal{N}^{b}\right) \tag{22}$$

where the $\mathcal{N}$ denotes the Gaussian noise variables, $\mathcal{N}^{\omega} \in \mathbb{R}^{q \times p}$ and $\mathcal{N}^{b} \in \mathbb{R}^{q}$. To deploy PSN in policy network, the perturbed policy network (PPN) and adaptive policy network (APN) are introduced. Importantly, PPN and APN have no process of updating parameters through gradient. Their parameters are only updated once per episode according to the parameters of policy network and the sampled noise and kept fixed until the next episode comes. Such mechanism can ensure the computational cheap. The relationships among the policy network, PPN and APN are shown in Fig. 4(a).

The PPN is used to make the decision according to the state during the training period. The action from the PPN is stored in experience replay and used to update the parameters of Q network. Unlike the traditional actor-critic DRL, the Q network is used to update the parameters of policy network instead of those of PPN.

Sampling PSN from the fixed noise distribution is not the desired way to deploy PSN to policy network since the impact of PSN on the results will seriously depend on the network structure, and the sensitivity of parameters to PSN will vary over the progress of training [24]. In this context, this paper uses APN to adaptively vary the noise distribution according to a certain metric. One alternative metric is that construct a distance measurement between the PPN and the policy network in the action space and adaptively increase or decrease the noise according to whether the parameter space noise is below or above threshold value. Specifically, it is assumed that the Gaussian noise is represented as $\mathcal{N}\left(0, \sigma^2\right)$, such metric can be defined as [24]:

$$\sigma_{k+1} = \begin{cases} \alpha\sigma_k, & \text{if } d\left(\pi, \tilde{\pi}\right) \le \delta \\ \dfrac{\sigma_k}{\alpha}, & \text{otherwise} \end{cases} \tag{23a}$$

$$d\left(\pi, \tilde{\pi}\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_s\left[\left(\pi(s)_i - \tilde{\pi}(s)_i\right)^2\right]} \tag{23b}$$

where the α is the scaling factor, the δ is a threshold value and N represents the dimensions of action.

The policy network is used to map the state s^g to action a^g , which plays the key role in the performance of the model. The deep neural network are useful extracting features form input to map to the desired output, which may strengthen the nonlinear mapping ability of policy network. However, simply increasing the depth of dense neural network may be invalid due to the original input information gradually disappear as the deepening of the network [25]. In order to overcome this problem, the architecture of policy network applied in proposed approach is shown in Fig. 4(b). As the Fig. 4(b) shows, the s^g is firstly input to h_1 dense neural network and output m_1, then, in order to mitigate the problem mentioned above, the input of h_2 is the concatenation of s^g and m_1 instead of m_1. Loop back and forth until the final a^g is output. This approach solves the problems skillfully by connecting the original input s^g with each hidden layer of the network to ensure that the s^g information can be retained to the maximum extent, thus making better use of the non-linear mapping ability of the deeper neural network. The more detailed information can refer to [25]. There are two reasons to apply D2RL in actor network: 1) PSN operates at the parameter level and D2RL builds a more complex and effective parametric connection relationship, which introduces greater potential for exploratory action than shallow dense architecture combined with PSN. 2) Compared with conditional shallow dense neural network architecture that applied in actor network [21]-[23], D2RL has stronger non-linear representational power to accelerate the establishment of mapping relationship from input state to optimal output decision making, which overcome the instability and slow convergence speed of training curves due to the introduction of noise.

The parameters of policy network can be optimized by ascent with the following gradient, $i \in [1, G]$:

$$\nabla J = \mathbb{E}_{s_k^{AG} \sim \mathcal{D},\, a_k^{AG} \sim \pi^{AG}\left(s_k^{AG}\right)}\left[\nabla \log\left(\pi^i\left(a_k^i \mid s_k^i\right)\right)\begin{pmatrix} Q^i\left(s_k^{AG},\, a_k^{AG}\right) \\ -\alpha \log \pi^i\left(a_k^i \mid s_k^i\right) \end{pmatrix}\right] \tag{24}$$

The policy target network can be updated by soft update mechanism, which can referred to [41].

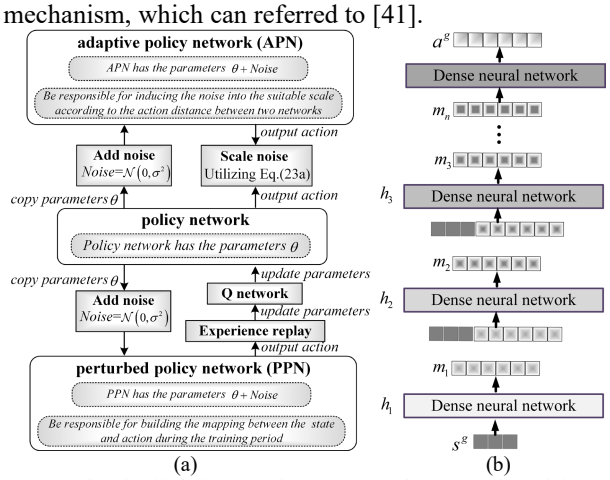


Fig. 4. The detailed framework of PSN and D2RL adopted in policy network: (a) The relationship between policy network, APN and PPN, and (b) The architecture of policy network.

The training process of MADRL-based approach are summarized in Algorithm 2. In Algorithm 2, each training

episode starts with the random selection of the electricity price of κ th day, t_{arr}^{g} and $SOC_{t_{arr}}^{g}$ of each agent at step 2 and 3, respectively. Then, the parameters of PPN are updated according to the parameters of policy network and sampled Gaussian noise to prepare for the start of the cycle at step 4. One cycle starts at $\min\{t_{arr}^{g}\}$ and ends at $\max\{t_{dep}^{g}\}$. From the step 6 to 8, the action is sampled from $\tilde{\pi}^{g}\left(a_t^{g} \mid s_t^{g}\right)$ where the $\tilde{\pi}^{g}$ belongs to the PPN instead of policy network. Then, the information of next time step can be obtained by interacting with the environment, and store the transition $\left(s_t^{g}, a_t^{g}, r_t^{g}, s_{t+1}^{g}\right)$ in experience replay. From the step 10 to 11, the parameters of critic and actor are updated based on the batch-sized sampled data. At the end of this episode, i.e., $t = \max\{t_{dep}^{g}\}$, the parameters of APN are updated in the similar way with the PPN while the noise distribution is updated based on the Eq. (23a).

Algorithm 2: Training process of MADRL-based approach

Inputs: The information mentioned in Eq. (15) of the G EVs.
Outputs: The charging/discharging power of the G EVs

1: **for** episode =1:M **do**
2: ▷ Randomly choose the electricity price of κ th day in range
3: ▷ Randomly choose t_{arr}^{g}, and $SOC_{t_{arr}}^{g}$ in range, where $g \in [1, G]$
4: ▷ Update the parameters of PPN
5: **for** $t = \min\{t_{arr}^{g}\}, \max\{t_{dep}^{g}\}$ **do** (Each agent executes in parallel)
6: ▷ Sample action a_t^{g} from $\tilde{\pi}^{g}\left(a_t^{g} \mid s_t^{g}\right)$
7: ▷ Enter s_t^{g} and a_t^{g} to environment to get r_t^{g} and s_{t+1}^{g}
8: ▷ Store transition $\left(s_t^{g}, a_t^{g}, r_t^{g}, s_{t+1}^{g}\right)$ in experience relay of g-th EV $\mathcal{O}^{g}$
9: **if** algorithm during update period **do**
10: ▷ Randomly sample batch-sized transitions from $\mathcal{O}^{g}$
11: ▷ Utilize the batch-sized sampled data to update parameters of critic and actor network by Eq. (20a) and (24)
12: **if** $t = \max\{t_{dep}^{g}\}$ **do**
13: ▷ Update the parameters of APN and excute Eq. (23a)
14: **end if**
15: **end if**
16: **end for**
17: **end for**

IV. Case Analysis

A. Case Study Setup

The electricity price for year 2017 zoom COMED and the basic load for the whole year 2017 of community residents utilized in this study are available online [43], [44]. The training set contains the first 200 days' data and the test set are from days 201–300 of 2017. For case studies, the scenario is considered one M=25 kVA distribution transformer serving G=4 residences where every residence has one EV, and each EV owner with RA_1.

The parameters of transformer LOL model are listed in Table I [11], [29]. For the detailed calculation process of LOL, readers can refer to [11]. The W_{ra} is assumed to be the same for all EV owners that are eager to have enough energy to reach the destination. The commuting behavior of EV owners are modeled as random variables [45]: the arrival time t_{arr}^{i} is sampled from $\{16, 17, 18, 19, 20, 21\}$ with equal probability to each one, t_{dep}^{i} is sampled from $\{7, 8, 9\}$ and SOC at arrival time SOC_{arr}^{i} follows normal distribution $\mathcal{N}\left(0.4, 0.1^2\right)$ bounded between 0.1 and 0.6. Four types of EV is considered, each for one residence [46]. The detailed parameters of EV owner's satisfaction model are shown in Table I. The detailed parameter setting of the LSTM-based forecasting model and MADRL-based control model are shown in Table II.

TABLE I
Parameters of Models Mentioned in Section II

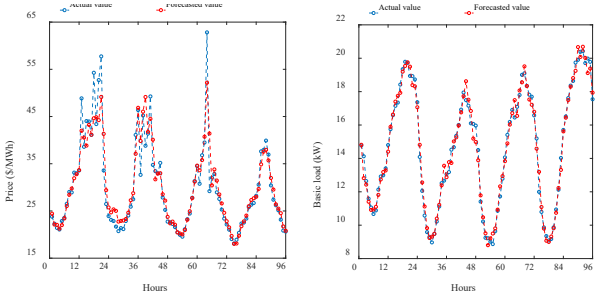
Model	Parameters
Transformer model	$n = 0.8$, $m = 0.8$, $\tau_{to} = 6.86$ hour, $\tau_{w} = 0.08$ hour, $\theta_{amb} = 40°\mathrm{C}$, $\Delta\theta_{to,r} = 53°\mathrm{C}$, $\Delta\theta_{hs,r} = 27°\mathrm{C}$, $R = 4.1$, $\mathrm{W}_{\mathrm{LOL}} = 2.5\times10^{4}\$$, $\mathrm{NLL} = 1.8\times10^{5}$ hours
EV owners' Satisfaction model	$\mathrm{W}_{\mathrm{ra}} = 2.5\$$, $t_{arr} \in \{16, 17, 18, 19, 20, 21\}$, $SOC_{t_{arr}} = \mathrm{clip}\left(\mathcal{N}\left(0.4, 0.1^2\right), 0.1, 0.6\right)$, $t_{dep} = \{7, 8, 9\}$, $\Upsilon = 0.0154436$

EV type	$E_{\max}$	$Power_{\max}$	$\mathcal{C}^{E}$
EV1: Leaf	24 kWh	6.3 kW	800$
EV2: BWM i3	18.8 kWh	5.4 kW	400$
EV3: Kia Soul	27 kWh	5.1 kW	200$
EV4: i-MIEV	16 kWh	3 kW	100$

TABLE II
The Parameters of the Proposed Approach

Model	Description	Value
LSTM-based	Learning rate	1e-4
	Batch sizes	256
	Training episodes	5000
	Optimizer	Adam
MADRL-based	Reward discount factor γ	0.95
	Temperature parameter α	0.1
	The capacity of experience relay $\mathcal{O}^{g}$	1e6
	Learning rate of actor	3e-4
	Learning rate of critic	3e-4
	Soft replacement τ	1e-3
	Batch sizes	256
	Training episodes	3e4
	Attention heads	4
	Threshold value δ	0.1
	Scaling factor α	1.01
	Optimizer	Adam

B. Performance of the LSTM-based Forecasting Model



(a) (b)

Fig. 5. Forecasted results: (a) Electricity price forecasting for days 201–204 of 2017, and (b) Basic load forecasting for days 201–204 of 2017.

The LSTM-based forecasting model is first trained utilizing historical data offline and then applied to forecast the future data. The forecasting results of electricity price and load demand on test set are shown in Fig. 5(a) and 5(b), respectively. It can be observed from Fig. 5(a) that the forecasting model can accurately predict the electricity price except for some values at the curve peak. Fig. 5(b) shows

that the load demand forecasted by the LSTM get very close to the real value, demonstrating the effectiveness of the forecasting model. To further evaluate the performance of the forecasting model, two commonly used metrics, mean absolute error (MAE) and mean absolute percentage error (MAPE) are utilized to assess the forecasted accuracy of the electricity price and load forecast [47]:

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|V_{true}^{i}-V_{forecast}^{i}\right| \tag{26a}$$

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|V_{true}^{i}-V_{forecast}^{i}\right|}{V_{true}^{i}}\times 100 \tag{26b}$$

where N denotes the amount of forecasted value used to calculate the metric; V_{true}^{i} is the i-th actual value, and $V_{forecast}^{i}$ represents the i-th forecasted value. The detailed values of MAE and MAPE of electricity price and basic load achieved by different approaches are summarized in Table III. Comparison approaches included Gated Recurrent Unit (GRU)-based model and back propagation neural network (BPNN)-based model. Comparing with the GRU and BPNN model, the MAEs and MAPEs in this work are lower, indicating the LSTM-based forecasting model in this paper can make reasonable and accurate electricity price and basic load predictions, which can benefit from the following decision-making process for coordinating different EVs charging/discharging.

TABLE III
MAE AND MAPE OF PREDICTION RESULT

Approach	Electricity price MAE / MAPE	Basic load MAE / MAPE
LSTM	2.37 / 9.32	0.323 / 2.49
GRU	2.39 / 9.42	0.329 / 2.51
BPNN	3.54 / 14.7	0.784 / 6.20

C. Performance of the MADRL-based approach

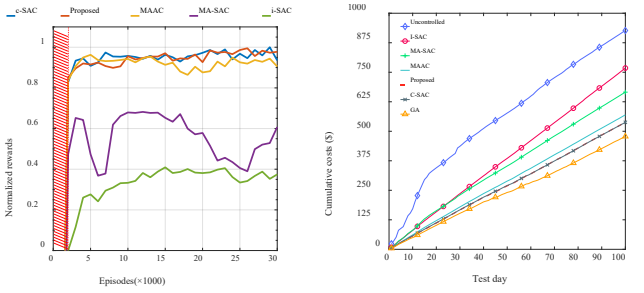
1) Description of Benchmarks

In order to evaluate the performance of proposed approach, five benchmarks, including uncontrolled approach; independent-SAC (i-SAC) approach based on the framework of ref [48]; multi-agent-SAC (MA-SAC) approach [49]; multi-actor-attention-critic (MAAC) approach [22]; centralized-SAC (c-SAC) approach based on the framework of ref [31]; and model-based centralized approach, i.e., genetic algorithm (GA) [50] are used for comparisons on a 4-EV coordinated charging system, which considers the transformer LOL and EV owners' satisfaction. The details of the whole benchmarks are as follows: ***Uncontrolled*****:** The EVs are charged immediately with the maximum charging power when they plugged. ***i-SAC*****:** i-SAC is an extension based on independent learner framework [48], which is independently optimized by its own reward function. In i-SAC, there are G independent agents to control G EVs. Each agent in i-SAC is trained independently to maximize its own reward according to the local observation, i.e. Eq. (15), which is different with the MA-SAC, MAAC and the proposed approach. ***MA-SAC*****:** MA-SAC is a MADRL approach proposed in [49] which has G critics (due to have the total G agents) where each critic has the information of concatenation of the whole agents' states and actions to build the coordinated relationship with each other. ***MAAC:*** MAAC is a MADRL approach [22], which has the attention layer to cooperative with each other. The architecture of actor network uses the traditional architecture mentioned in [19]-[23], which is different from the proposed approach. ***c-SAC*****:** c-SAC is an extension based on one to many RL control frameworks [31], where all the EVs are controlled by a central controller. The c-SAC has the global information in both training and execution process, which is different with i-SAC, MA-SAC, MAAC and proposed approach. ***GA*****:** GA [50] is a meta-heuristic swarm intelligent technology, which has the advantage of parallel search capability, wide global search, and probabilistic transition rules to guide its search direction.

Therein, GA optimizes the model when the electricity price and commuting behavior of EV owners are known. The MA-SAC, i-SAC, MAAC, c-SAC, and proposed approach only utilize the predicted electricity price, and unknown EV owners' commuting behavior before.

2) Simulation Results

The training process of different RL-based approaches on training set is shown in Fig. 6(a), where the reward is an average of every 1000 values. The red shaded part in Fig. 6(a) represents the unregulated exploration process of agents, corresponding to 6-8 steps of algorithm 2. The reward curve to the right of the red shaded part indicates that the network parameters are beginning to be optimized, and we normalize the reward curve in the optimization process to facilitate comparison between different approaches. It can be observed from the figure that at the beginning, no approaches can make good decisions to obtain high cumulative reward. With the ongoing of the training, these approaches gradually learn the control strategy to achieve high reward and ultimately converge. Fig. 6(b) shows the cumulative costs of different approaches over the 100 test days. The detailed comparative results achieved by various approaches on the test set are summarized in Table IV. When uncontrolled approach is applied, unmanaged charging behavior results in maximum cost in all approaches. Compared with uncontrolled approach, the i-SAC approach can reduce the cumulative cost to some extent. However, since each agents is trained independently in the i-SAC approach, such a way has the potential to lead to lack of coordination among agents, and thus its cumulative cost is larger than MA-SAC approach that can learn a coordinated charging strategy during the centralized training stage. Obviously, the attention layer is utilized in the MAAC approach further enhance the coordination between different agents, thus it achieves better performance than the MA-SAC approach. The proposed approach uses D2RL to eliminate the instability and inefficiency caused by noise during PSN exploration, while introducing richer parameter connections to form an effective exploration mechanism in combination with PSN. The proposed approach utilizes this effective exploration mechanism to achieve the smaller cumulative cost than MAAC approach. The c-SAC approach for decision making based on global information has the visible gap compared with the MAAC approach but it only has a very small advantage over proposed approach. The global information is difficult to obtain since the collection of commuting behavior of other EV owners may comprise the privacy, and thus it is particularly important for EV charging management that the local information-based proposed approach has similar performance to the global information-based c-SAC approach. In the optimization of GA, we assume that the uncertain variables are known in advance, so that the EVs coordinated charging management can be formulated as a deterministic optimization problem and solved by GA. However, it cannot be deployed in the realistic scenario due to the existence of randomness. In the table IV, the GA has the negative LOL cost means that the EV owners can obtain the benefit from reducing the LOL cost. This mechanism is reasonable due to the EV charging behavior is controlled to minimize the transformer LOL at the expense of EV owners' benefits [8].



(a) (b)

Fig. 6. The simulation results on the training and test set. (a) Average rewards during the training process, and (b) The total four EVs' cumulative costs of the different approaches over the 100 test days.

TABLE IV
THE DETAILED DATA IN FIG. 6(B)

	Uncontrolled	i-SAC	MA-SAC	MAAC	Proposed	c-SAC	GA
Cumulative cost ($)	927	779	666	570	538	538	478
LOL cost ($)	281	23.0	58.5	12.1	4.82	7.18	-7.4
Range anxiety cost ($)	0	96.2	211	185	188	188	213
Charging cost ($)	230	203	160	120	98.2	94.2	-21
Degradation cost ($)	415	457	237	252	247	249	293

In this paper, the proposed approach aims to optimize four objectives simultaneously: 1) minimizing the transformer LOL; 2) minimizing the charging cost of EV owners; 3) minimizing the EV owners' range anxiety; 4) minimizing the cost due to the battery degradation. The detailed coordinated charging results over 3 consecutive days are shown in Fig. 7. The blue regions in both Fig. 7(a) and Fig. 7(b) represent the time when EVs leave home. In Fig. 7(a), the arrows denotes the time of four EVs when arriving home, the blue line represents $\theta_{hs,t}$ which is mentioned in Eq. (15), the red line represents the real-time hourly electricity price, and the bar denotes the charging/discharging power of EV. In order to have clear comparison among the four EVs in Fig. 7(a), the real-time charging/discharging power of the four EVs are normalized based on the maximum charging power.

It can be observed from Fig. 7(a) that the proposed approach learns to charge when the electricity price and temperature are low and discharge when the price and temperature are high to minimize the transformer LOL and charging cost, see from time 1 to 14 and time 25 to 39 for example. It can be observed that time 11 and 31 have similar high price and temperature but the difference between them is that the time 11 are at the end of an episode and will consider the range anxiety more, while the latter is at the beginning and will consider transformer LOL and charging cost more than range anxiety. Since the discharging operation will aggravate the degradation of the battery [32], the revenue the owner obtains from discharging the battery is in conflict with the cost caused by the degradation of the battery. In this work, the discharge preferences of EV owners are differentiated by setting the C^E value, larger C^E will result in larger battery cost due to the battery degradation. Owing to the coefficient C^E of EV 1 is larger than other EVs, the agent of EV 1 tend to discharge less power than EV 2, 3, and 4. This phenomenon can be observed in time 4, 25, 26, 29 and 30. However, having a high cost of battery degradation does not mean that EV will not benefit from large discharging actions, except that more benefits are needed to offset the cost of battery degradation due to discharging, time 3 is an example. At time 3, the very high discharging of EVs not only good for transformer, but also the EV owners (even to EV1). Therefore, if the benefits of discharging are considerable, the EV with high battery degradation cost will also perform large discharging actions to obtain greater benefits. The above-mentioned phenomenon indicate that the proposed approach can make flexible decisions according to the actual situation, in order to maximize the overall benefits.

In Fig. 7(b), the bar represents the SOC, the deep blue dotted line and the blue line is the winding hottest-spot temperature after using the proposed approach and uncontrolled approach to manage the four EVs charging, respectively. The temperatures represented by these two lines are different from the $\theta_{hs,t}$ mentioned in Eq. (15) in that they are the hottest-spot temperature after being loaded by $load_t^{total}$, while $\theta_{hs,t}$ only considers the $load_t^{basic}$. As shown in Fig. 7(b), the SOC of four EVs can achieve the goal of fully charging before leaving home. Comparing the deep blue dotted line and blue line can be observed that the proposed approach can well-coordinate the four EVs charging to cut the peak of the hottest-spot temperature. This is meaningful for reducing the transformer LOL. In order to have a clear description between hottest-spot temperature and transformer LOL, the LOL% difference when the temperature difference is fixed at 5°C are shown in Fig. 8. As the Fig. 8 shows, the first dot denotes the value of LOL% at 75°C hottest-spot temperature minus that of LOL% at 70°C. The meaning of the difference is the lifetime damage to the transformer caused by the temperature rise of 5°C on the basis of 70°C. The trend of the line in Fig. 8 shows that for the transformer in the peak of the hottest-spot temperature, every increase in peak temperature value will cause huge damage to the transformer than before and as such peak clipping is a good way to prolong transformer life. Back to the Fig. 7(b), from the figure can be observed that the highest temperature peak under the uncontrolled approach is 141.38°C at time 27, while the temperature of the proposed approach is 98.29°C at the same time. This is because the proposed approach choose to discharge at time period 25 to 27 to reduce the temperature, and at time period 33 to 38 to charge the battery to reduce the EV owners' range anxiety. This operation moves the charging area from time 25 to 27 (uncontrolled approach charging area) to time 33 to 38 (proposed approach charging area), which not only significantly reduces the transformer LOL, but also minimize the EV owner's range anxiety. The above mentioned results demonstrate that the proposed approach can simultaneously reduce the transformer LOL and EV owner's dissatisfaction.

Further tests are carried out to evaluate the impact of the range anxiety function on the performance of the proposed control approach. Four cases with the same commuting behaviors are considered in this test: 1) case 1, where four EV owners are included and RA_1 is utilized to capture the range anxiety effect of all EV owners; 2) case 2, where RA_2 is used to represent the range anxiety effect; 3) case 3, where RA_3 is used to capture the range anxiety effect of all EV owners; 4) case 4, where RA_1, RA_2 and RA_3 are used to represent the range anxiety effect of EV1 and EV4, EV2, and EV3, respectively. The cumulative costs achieved by different approaches under the four cases are shown in Fig. 9. It can be observed from the figure that the optimization results are sensitive to the selection of the range anxiety function. Since the RA_1 penalty value is higher than RA_2 and RA_3, the battery energy is more abundant than

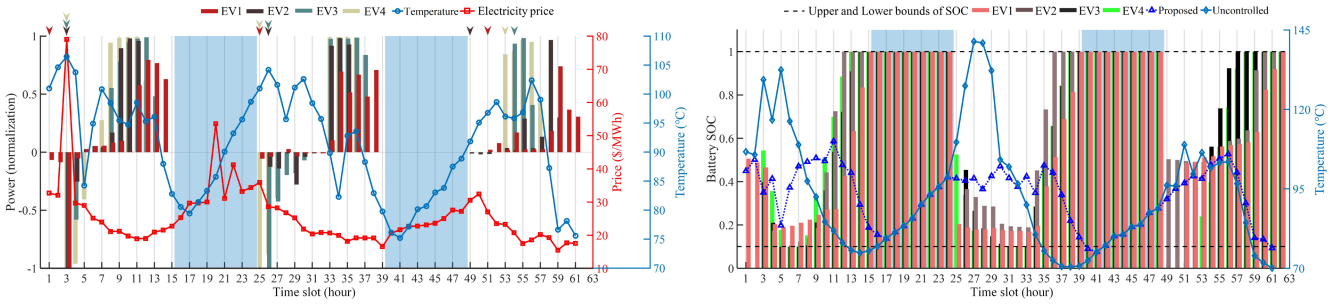


Fig. 7. Detailed coordinated charging results over 3 consecutive days. (a) Including four EVs' charging/discharging scheduling of the proposed approach, hourly electricity price, and the θ_{hs} before EV loading over 3 consecutive days, and (b) Including four EVs' remaining SOC, the time varying curve of θ_{hs} using the proposed and the uncontrolled approach over 3 consecutive days.

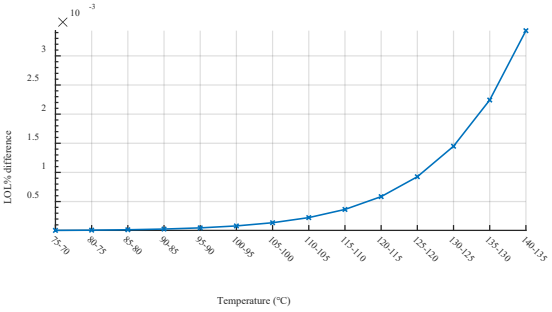


Fig. 8. LOL difference when the temperature difference is fixed at 5°C.

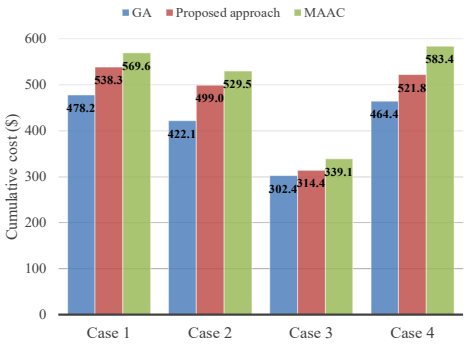


Fig. 9. Cumulative cost of three approaches in four cases.

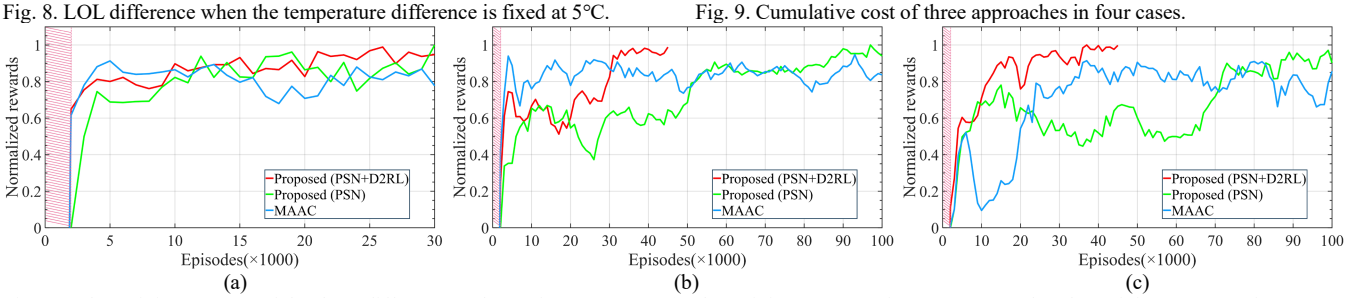


Fig. 10. The training process of the three different numbers of agent cases. (a) The training process of 4-agents case, (b) The training process of 8-agents case, and (c) The training process of 12-agents case.

others, and thus the transformer LOL, battery degradation and charging cost may have quite different from that of other cases, which leads to the cumulative costs achieved by different approaches under case 1 are normally higher than that obtained under case 2, case 3 and case 4. Case 3 achieves the least cumulative cost, but the battery is typically not fully charged under this case due to the characteristics of RARR of RA_3. The proposed approach can always achieve control performance that is better than MAAC approach and gets close to that obtained by GA approach under all cases, demonstrating the effectiveness of the proposed approach.

To further investigate the effectiveness of the proposed approach, the scalability comparisons between proposed approach and MAAC are shown in Fig. 10, which are the training processes of 4-agents, 8-agents, and 12-agents cases. The experimental parameters of 4-agents case have been summarized in table I. The 8-agents and 12-agents cases are the extensions of 4-agents case. Specifically, the 8 EVs in 8-agents case and 12EVs in 12-agents case consist of the double and triple EVs in 4-agents case, respectively. Similarly, the basic load, transformer capacity and W_{LOL} change proportionally to denote the increase of EV owners. As the Fig. 10 shown, from the 4-agents case to 12-agents case, the proposed approaches have the better performance than that of the MAAC. The reason for this phenomenon is that unlike single-agent environments, the action space of multi-agent environments increases exponentially with the number of agents increasing. Such a mechanism causes any change in the number of agents to affect the action space in an exponential manner. Due to the PSN offers the strong exploration ability to avoid the prematurely converge on the optimization of large-scale action space, the green reward curves have the better performance than the blue reward curves in these three cases. However, due to the lack of nonlinear representational power of traditional shallow dense architecture of actor network, the convergence speed and the stability of training are not good enough. In view of this, the D2RL is applied in actor network. Comparing the red and green reward cure can be seen that the former converges faster and more stable than the latter, especially this gap gradually increases with the increase of the number of agents. The results prove that D2RL provides the more complex and effective parametric connection relationship compares with shallow dense architecture of actor network for PSN operated at the parameter level to result in achieving the better control performance.

V. Conclusions

This paper proposes a MADRL enabled decentralized approach for the optimization of LOL of transformer considering the requirements of EV owners. LSTM is first utilized to capture the uncertainties of the electricity price and load demand. Then the coordinated scheduling of multiple EVs is cast to a Markov game, which is solved by the proposed MADRL approach features centralized training and decentralized execution. The centralized training procedure helps the formulation of a coordinated control strategy, which is further enhanced by the attention mechanism. In addition, the PSN and D2RL are introduced to overcome premature convergence, training instability and inefficiency due to the large action space of multi-agent scenario. Since only local information are utilized during

execution stage, the privacy of EV owners are preserved, the related communication cost are reduced and the single-point failure can be avoided. Comparative results demonstrate that the critic network processes the entire EV information by using the attention mechanism and effectively guides the generation of coordinated strategies among actor networks; the actor network utilizes the combination of PSN and D2RL to achieve a better, faster, and smoother training effect in the training phase, and the experimental results on the test set similarly verify its effectiveness.

VI. References

[1] Q. Gong, S. Midlam-Mohler, E. Serra *et al.*, "PEV charging control considering transformer life and experimental validation of a 25 kVA distribution transformer," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 648-656, Mar. 2015.

[2] S. S. Karimi Madahi, H. Nafisi, H. Askarian Abyaneh *et al.*, "Co-Optimization of energy losses and transformer operating costs based on smart charging algorithm for plug-In electric vehicle parking lots," *IEEE Transactions on Transportation Electrification*, vol. 7, no. 2, pp. 527-541, June 2021.

[3] N. Jabalameli and A. Ghosh, "Online centralized coordination of charging and phase switching of PEVs in unbalanced LV networks with high PV penetrations," *IEEE Systems Journal*, vol. 15, no. 1, pp. 1015-1025, Mar. 2021.

[4] M. R. Sarker, M. A. Ortega-Vazquez and D. S. Kirschen, "Optimal coordination and scheduling of demand response via monetary incentives", *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1341-1352, May 2015.

[5] D. J. Olsen, M. R. Sarker and M. A. Ortega-Vazquez, "Optimal penetration of home energy management systems in distribution networks considering transformer aging," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3330-3340, July 2018.

[6] M. Yilmaz and P. T. Krein, "Review of battery charger topologies, charging power levels, and infrastructure for plug-In electric and hybrid vehicles," *IEEE Transactions on Power Electronics*, vol. 28, no. 5, pp. 2151-2169, May 2013.

[7] Y. Shang, W. Wu, X. Huai *et al.*, "Loss of life estimation of distribution transformers considering corrupted AMI data recovery and field verification," *IEEE Transactions on Power Delivery*, vol. 36, no. 1, pp. 180-190, Feb. 2021.

[8] M. R. Sarker, D. J. Olsen and M. A. Ortega-Vazquez, "Co-Optimization of distribution transformer aging and energy arbitrage using electric vehicles," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2712-2722, Nov. 2017.

[9] O. Beaude, S. Lasaulce, M. Hennebel *et al.*, "Reducing the impact of EV charging operations on the distribution network," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2666-2679, Nov. 2016.

[10] X. Zhang, E. Gockenbach, V. Wasserberg *et al.*, "Estimation of the lifetime of the electrical components in distribution networks," *IEEE Transactions on Power Delivery*, vol. 22, no. 1, pp. 515-522, Jan. 2007.

[11] IEEE Std C57.91-2011: "IEEE guide for loading mineral-oil-immersed transformers and step-voltage regulators," 2011.

[12] Q. Gong, S. Midlam-Mohler, V. Marano *et al.*, "Study of PEV charging on residential distribution transformer life," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 404-412, Mar. 2012.

[13] Z. Ma, D. Callaway and I. Hiskens, Control and Optimization Methods for Electric Smart Grids, New York, NY, USA:Springer, 2012.

[14] M. Soleimani and M. Kezunovic, "Mitigating transformer loss of life and reducing the hazard of failure by the smart EV charging," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5974-5983, Sept.-Oct. 2020.

[15] Q. Gong, S. Midlam-Mohler, V. Marano *et al.*, "Distribution of PEV charging resources to balance transformer life and customer satisfaction," in *Proceedings of IEEE Int. Elect. Veh. Conf. (IEVC)*, Greenville, SC, USA, Mar. 2012, pp. 1–7.

[16] Siobhan Powell, Emre Can Kara, Raffi Sevlian *et al.*, "Controlled workplace charging of electric vehicles: The impact of rate schedules on transformer aging", *Applied Energy*, vol. 276, 2020.

[17] Y. Liang, F. Liu and S. Mei, "Distributed real-time economic dispatch in smart grids: a state-based potential game approach," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4194-4208, Sept. 2018.

[18] F. L. D. Silva, C. E. H. Nishida, D. M. Roijers *et al.*, "Coordination of electric vehicle charging through multiagent reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2347-2356, May 2020.

[19] B. Gu, X. Yang, Z. Lin *et al.*, "Multiagent actor-critic network-based incentive mechanism for mobile crowdsensing in industrial systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 6182-6191, Sept. 2021.

[20] K. Zhang, J. Cao and Y. Zhang, "Adaptive digital twin and multi-agent deep reinforcement learning for vehicular edge computing and networks," *IEEE Transactions on Industrial Informatics*, doi: 10.1109/TII.2021.3088407.

[21] D. Cao, W. Hu, J. Zhao *et al.*, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120-4123, Sept. 2020.

[22] L. Yu, Y. Sun, Z. Xu *et al.,* "Multi-agent deep reinforcement learning for HVAC control in commercial buildings," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 407-419, Jan. 2021

[23] R. Lu, Y. Li, Y. Li *et al.*, "Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management", *Applied Energy*, vol. 276, Oct. 2020.

[24] M. Plappert, R. Houthooft, P. Dhariwal *et al.*, "Parameter space noise for exploration," in *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[25] S. Sinha, H. Bharadhwaj, A. Srinivas *et al.*, "D2RL: Deep Dense Architectures in Reinforcement Learning," arXiv:2010.09163, 2021.

[26] M. Fortunato, M. G. Azar, B. Piot *et al.*, "Noisy networks for exploration," in *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[27] D. Cao, W. Hu, J. Zhao *et al.*, "Reinforcement learning and Its applications in modern power and energy systems: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029-1042, November 2020.

[28] Z. Wan, H. Li, H. He *et al.*, "Model-free real-Time EV charging scheduling based on deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5246-5257, Sept. 2019.

[29] R. Vicini, O. Micheloud, H. Kumar *et al.*, "Transformer and home energy management systems to lessen electrical vehicle impact on the grid," *IET Generation, Transmission & Distribution*, vol. 6, no. 12, pp. 1202-1208, December 2012.

[30] N. Chen, M. Wang, N. Zhang *et al.*, "Energy and information management of electric vehicular network: a survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 967-997, Second quarter 2020.

[31] H. Li, Z. Wan and H. He, "Real-time residential demand response," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4144-4154, Sept. 2020.

[32] M. A. Ortega-Vazquez, "Optimal scheduling of electric vehicle charging and vehicle-to-grid services at household level including battery degradation and price uncertainty," *IET Generation, Transmission & Distribution*, vol. 8, no. 6, pp. 1007-1016, June 2014.

[33] M. L. Littman, "Markov games as a framework for multiagent reinforcement learning," *Machine Learning Proceedings* 1994, pp. 157–163. Elsevier, 1994.

[34] S. Zhou, L. Zhou, M. Mao *et al.*, "An optimized heterogeneous structure LSTM network for electricity price forecasting," *IEEE Access*, vol. 7, pp. 108161-108173, 2019.

[35] W. Kong, Z. Y. Dong, Y. Jia, *et al.*, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, Jan. 2019.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735-1780, 1997.

[37] Felix A. Gers, Nicol N. Schraudolph, J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115-143, 2003.

[38] S. Iqbal, and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning." in *International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 2961–2970.

[39] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, USA, Dec. 2017, pp. 5998–6008.

[40] A. Krizhevsky, I. Sutskever and Geoffrey E. Hinton "ImageNet classification with deep convolutional neural networks," *Communications of The ACM*, vol. 60, no. 6, 2017, pp. 84–90.

[41] T. Haarnoja, A. Zhou, P. Abbeel *et al.*, "Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of International Conference on Machine Learning (ICML)*, Stockholm, Sweden, July, 2018.

[42] Ian Osband, Benjamin Van Roy, Daniel J. Russo *et al.,* "Deep exploration via randomized value functions," *Journal of Machine Learning Research*, vol. 20, no. 124, pp. 1–62, 2019.

[43] "Historical hourly electricity price from PJM", 2017. [Online]. Available: https://www.pjm.com/.

[44] "Historical hourly load data", 2017. [Online]. Available: https://www.nationalgridus.com/.

[45] S. Li, W. Hu, D. Cao *et al.,* "Electric vehicle charging management based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, doi: 10.35833/MPCE.2020.000460.

[46] K. Chaudhari, N. K. Kandasamy, A. Krishnan *et al.,* "Agent-based aggregated behavior modeling for electric vehicle charging load," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 856-868, Feb. 2019.

[47] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Applied Energy*, vol. 236, pp. 937-949, May 2019.

[48] X. Xu, Y. Jia, Y. Xu, *et al.,* "A multi-agent reinforcement learning-based data-Driven method for home energy management," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201-3211, July 2020.

[49] D. Cao, J. Zhao, W. Hu *et al.,* "Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of PVs," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 4137-4150, Sept. 2021.

[50] C. Chiang, "Improved genetic algorithm for power economic dispatch of units with valve-point effects and multiple fuels," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 1690-1699, Nov. 2005.