A Reinforcement Learning-based Assignment Scheme for EVs to Charging Stations

Mohammad Aljaidi*, Nauman Aslam*, Xiaomin Chen*, Omprakash Kaiwartya†,
Yousef Ali Al-Gumaei*, Muhammad Khalid‡
*Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK
†School of Science and Technology, Nottingham Trent University, UK
‡Department of Computer Science and Technology, University of Hull, UK
Email:*{mohammad.jaidi,nauman.aslam,xiaomin.chen,yousef.al-gumaei)@northumbria.ac.uk,
†omprakash.kaiwartya@ntu.ac.uk, ‡m.khalid@hull.ac.uk

***Abstract*—Due to recent developments in electric mobility, public charging infrastructure will be essential for modern transportation systems. As the number of electric vehicles (EVs) increases, the public charging infrastructure needs to adopt efficient charging practices. A key challenge is the assignment of EVs to charging stations (CSs) in an energy efficient manner. In this paper, a Reinforcement Learning (RL)-based EV Assignment Scheme (RL-EVAS) is proposed to solve the problem of assigning EV to the optimal CS in urban environments, aiming at minimizing the total cost of charging EVs and reducing the overload on Electrical Grids (EGs). Travelling cost that is resulted from the movement of EV to CS, and the charging cost at CS are considered. Moreover, the EV's Battery State of Charge (SoC) is taken into account in the proposed scheme. The proposed RL-EVAS approach will approximate the solution by finding an optimal policy function in the sense of maximizing the expected value of the total reward over all successive steps using Q-learning algorithm, based on the Temporal Difference (TD) learning and Bellman expectation equation. Finally, the numerous simulation results illustrate that the proposed scheme can significantly reduce the total energy cost of EVs compared to various case studies and greedy algorithm, and also demonstrate its behavioural adaptation to any environmental conditions.**

***Index Terms*—Electric vehicle assignment, charging station, Q-learning, temporal difference, Bellman expectation equation, energy consumption, energy cost, electrical grids.**

I. Introduction

Being one of the fastest growing sources of energy demand and carbon dioxide (CO_2) emissions, the transportation sector is under great pressure to be decarbonized through deploying EVs [1]. At the same time, various parameters, such as the technological innovation in battery efficiency and electric drivetrain, led to dramatically increase the EVs penetration in recent years in urban environments [2]. Charging EVs at home is an alternative solution for the EV users, however, it takes too much time (6 to 8 hours). Therefore, establishing a conveniently available fast-CSs is essential to increase the satisfaction of EV users, and alleviating peak charging loads. Using fast-CSs can charge EV batteries at least 12 times faster [3]. The adoption rate of EVs mainly depends on the presence of wide range of CSs in metropolitan areas to help EV users to charge their vehicles during their journeys. Furthermore, the selection of best fast-CSs for EVs is an important factor that affects not only the spread of EVs but also minimize the total energy consumption of EVs to reach CSs and increase the sustainability of transportation, knowing that the energy demand in the transportation sector will increase by 54% until 2035 [4].

In recent years, more attention has been paid to suggesting learning-based techniques for finding the optimal CSs for EVs in metropolitan environments. Several algorithms and technique have been utilized to solve this problem. Machine learning is a sub-field of artificial intelligence (AI) that provides systems with the ability to automatically learn from experience without being explicitly programmed and finally solve the most complex problems [5]. RL technique is one of the machine learning categories that has been utilized recently in several studies to solve the problem of assigning EV to optimal CS. In [6], an optimal CS selection algorithm based model-free RL technique has been proposed, to minimize the total travel time of EVs charging demands from origin to CS using the selected optimal route, taking into account unknown future demands and dynamically changing traffic conditions. The CS selection was formulated as a Markov decision process (MDP) model with unknown transition probability.

A congestion control in CSs allocation with Q-learning has been introduced in [7]. In this paper, the authors have considered the travel time and the queuing time in CSs in their model, to build joint-resource congestion game which utilises the interaction between the EVs and resources. Q-learning algorithm has been applied to the model in order to solved the problem. In [8], a deep RL (DRL) based approach for EV charging navigation has been proposed to adaptively learn the optimal technique for EV charging navigation without any prior knowledge of uncertainties. Feature states have been extracted from stochastic data. The authors investigated the DRL-based charging navigation from a single perspective of EV owner, considering randomness in smart grid and intelligent transportation data. To the best of our knowledge, the existing literature did not take into account the variation in the charging rate between CSs, in order to assign EVs to the optimal CSs. Our proposed scheme is unique in that, as we take into account the differences in charging rate among all available CSs, the charging cost at CSs, the travelling cost of EV to reach CS, and also the capability of the CSs. We argue that all of these metrics have a significant influence on the decision of assignment EVs to CSs in urban areas.

The main contributions of the present paper are summarized as follows:

- A RL-scheme for assignment of EVs to the optimal CSs in metropolitan environments is proposed in this paper. The proposed scheme considers the energy consumption cost that is resulted from the movement of EV towards CS (travelling cost), and the total expected cost to fully charge EV at CS (total energy cost). An EV's battery SoC in the process of finding the optimal CS is also taken into account.
- Q-learning algorithm has been utilized to solve this problem based on maximizing the cumulative reward of the EV during learning process by reducing the total cost of charging EVs.
- As a results of applying our proposed scheme, we minimize the load on the overwhelmed EGs, by assuming different rewards for CSs in the study area. The reward at each CS is determined based on the electricity price offered by EGs' suppliers to CSs, these prices vary according to the load and locations of EGs.

The rest of the paper is organized as follows. EV assignment problem formulation and optimization model are presented in Section II. Then, a RL based approach is proposed in Section III. Section IV shows the numerical results of our proposed approach. Finally, Section V draws a conclusion.

II. EV Assignment Problem

In this section, the assignment and optimization problem will be presented. The notations used in this paper are listed in Nomenclature list. In addition, parameters and variables are explained in the where they are first used.

Nomenclature

Sets and Index

$\mathcal{A}$	Set of actions.
$\mathcal{M}$	Set of CSs.
$\mathcal{S}$	Set of states.
j	index of CSs.

Parameters and Variables

α, γ	The learning rate and discount factor, respectively.
μ_{ev}, μ_j	The longitude of EV and CS_j, respectively.
Φ	A threshold value that restricts the maximum amount of energy consumption that EV can consume to reach CS
$\pi^*(s)$	The optimal policy.
ψ_{ev}, ψ_j	The latitude of EV and CS_j, respectively.
φ_j	The charging rate at CS_j.
$\vartheta_{ev,j}$	The energy consumption cost resulted from the movement of EV towards CS.
$\xi_{ev,j}$	The total expected energy to fully charge EV at CS, except the travelling cost.
ζ_{ev}	The EV battery capacity.
a, a'	The action that the agent takes in the current state, and target state, respectively.
$C_{ev,j}$	The overall cost of charging EV.
$d_{ev,j}$	The distance that EV travels to reach CS.
$E_{ev,j}$	The overall energy.
M	The number of CSs in the study area.
$Q^*(s,a)$	The optimal state-action value function.
$Q(s,a)$	A state-action value function.
r	The immediate reward.
Rw_j	The reward that is associated with CS_j.
s, s'	The current and target states, respectively.
$U_{ev,j}$	The accumulative reward for EV with CS_j.
V	The total number of actions that EV takes to arrive CS_j.
$x_{ev,j}$	A binary decision variable shows that the EV selects CS_j for charging.

A. Problem Formulation

The problem has been formulated as shown in the following sections:

1) EV: The EV is represented by a single agent that moves in the environment trying to find the optimal CS considering the possible actions and reward at each state. The EV has one attribute; (p^{ev}), where p^{ev} is the position (coordinates) of EV.

2) CSs: Define the CS set as $\mathcal{M} = \{1, ..., u, ..., M\}$. The cardinality of $\mathcal{M}$ is M, i.e., there are M CSs in the investigated area. CS_j in $\mathcal{M}$ has two attributes; b^u, r^u, where b^u, and r^u are the position and reward of the CS. The reward at CS_j, depends on the charging rate of CS_j.

3) Cost-based EV assignment: The proposed strategy uses a RL technique to assign EV to the best CS based on minimizing the total cost of charging EVs, i.e, $C_{ev,j}$. To calculate the total expected cost of charging EV, two factors should be investigated: the energy consumption cost that is resulted from the travelling of EV towards CS, i.e., $\vartheta_{ev,j}$, and the total expected energy to fully charge EV at CS, i.e., $\xi_{ev,j}$, considering the charging rate at each CS, which is usually determined based on the electricity price offered by EGs to the CS owners. The electricity price that offered by EGs to CSs is different due to the load and location of each EG. CSs connected to the same EG have the same electricity price, and therefore the same charging rate. The overall energy, i.e, $E_{ev,j}$, can be calculated as follows:

$$E_{ev,j} = \vartheta_{ev,j} + \xi_{ev,j} \tag{1}$$

To calculate $\vartheta_{ev,j}$, we need to calculate the amount of energy that the EV consumes per km to reach CS, i.e., δ_{ev} [9], [10], and also calculate the total distance that the EV travels towards CS, i.e., $d_{ev,j}$. In this work, we assume that δ_{ev} is 0.16 kWh/km [11], [12] The following illustrates how $d_{ev,j}$, $\vartheta_{ev,j}$, and $\xi_{ev,j}$ are calculated:

$$d_{ev,j} = \sqrt{(\psi_{ev} - \psi_j)^2 + (\mu_{ev} - \mu_j)^2} \tag{2}$$

where ψ_{ev}, ψ_j are the latitude of EV and CS_j, and μ_{ev}, μ_j are longitude, respectively.

$$\vartheta_{ev,j} = \delta_{ev} * d_{ev,j} \tag{3}$$

where δ_{ev} is the amount of energy that the EV consumes per km to reach CS. Eq. 4 shows how the total energy to fully charge EVs is calculated, except the travelling cost.

$$\xi_{ev,j} = (\zeta_{ev} - SoC) \tag{4}$$

where ζ_{ev} denotes the capacity of the battery, and SoC is the battery state of charge. The total cost of charging EV can be calculated as follows:

$$C_{ev,j} = E_{ev,j} \times \varphi_j \tag{5}$$

where φ_j represents the charging rate at CS_j.

B. Optimization Problem

The corresponding optimization problem of our proposed approach can be written as:

$$\min_{X} \sum_{j=1}^{M} C_{ev,j}\ x_{ev,j} \tag{6}$$

$$s.t. \quad \sum_{j=1}^{M} x_{ev,j} = 1 \tag{7}$$

$$x_{ev,j} \in \{0,1\}, \qquad \forall j \tag{8}$$

$$\vartheta_{ev,j} < \Phi, \qquad \forall j \tag{9}$$

where the variable $x_{ev,j}$ is a decision variable with binary values {0,1} as shown in Eq. (8), to indicate whether the CSj is selected by the EV or not, $x_{ev,j}$ is equal to 1 if the j is selected, otherwise it is equal to 0. Constraint (7) restricts that only one CS is selected as destination. Constraint (9) shows that the total amount of energy consumption that EV needs to reach CS should not reach to a certain threshold Φ in order to maintain the EV battery's SoC.

III. Reinforcement Learning Approach

Typically, RL techniques can be processed under two categories: off-policy and on-policy [13]. In particular, an off-policy learning method (e.g., Q-learning) earns an optimal target policy independent of the behavior policy used during exploration process as long as the different states are explored enough times. Whereas on-policy learning method finds the optimal policy taking into account the actual actions taken over the exploration process, which means that the target policy is the same as the behavior policy used in exploration process. Q-Learning technique will be used in this work to addresses the problem of assignment EV to CS.

A. Q-Learning-based EV Assignment

In this section, a RL technique is employed to solve our optimization problem (6-9), based on Q-Learning Algorithm technique. Q-learning, an incremental technique for dynamic programming, is appropriate for solving such kind of problems. Q-learning is an agent-based technique in which the AI agent interacts with its environment and adapts its actions based on rewards or penalties received in response to its actions [14]. Mainly, there are three basic elements in the algorithm: environment, state, and action. We will introduce the algorithm after setting the elements.

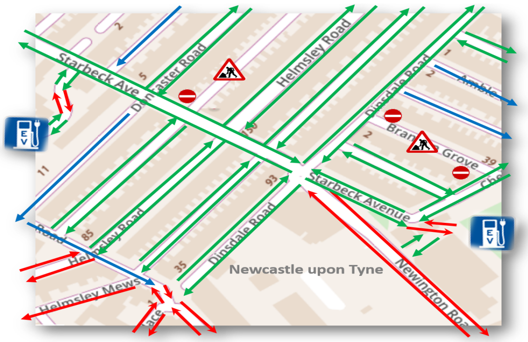

Fig. 1. Sample from the urban area of Newcastle upon Tyne city, UK

1) Environment, State, and Action Set: The environment is an essential element in Q-learning, in which the AI agent selects its actions according to corresponding rewards. In our scenario, the environment should involve roads between EVs and those available CSs. Fig. 1 shows part of the Newcastle upon Tyne, UK. The directions (actions) the EV is allowed to take in the study area, are represented by three colors of arrows. The green and blue arrows show that the EV can move in only four directions: South, North, East, and West. The difference between the green and blue arrows is that the green arrows show that an EV can move in both directions, while the blue arrows show that an EV can only move in one direction. The red arrows indicate that an EV can move in the other four directions: South-East, South-West, North-East, and North-West and also in both directions. Road works signs indicate that these streets are closed and cannot be used by vehicles, so the EV user needs to find other streets to reach CS. The CSs are distributed in fixed locations in the study area as shown in Fig. 1.

The state set in our scenario can be denoted by $\mathcal{S}$, and defined as following:

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)\} \tag{10}$$

where x and y represent the position (coordinate) of the state that an agent can visit during its journey to the destination. The cumulative reward is calculated based on the actions that have been taken by an EV in the environment, and can be calculated as follows:

$$U_{ev,j} = Rw_j - \sum_{i=1}^{V} r_i \tag{11}$$

where $U_{ev,j}$ represents the accumulative reward for EV with CS_j, Rw_j denotes the reward that is associated with CS_j, V represents the total number of actions that EV takes to arrive CS_j, and r_i is the immediate reward that EV gets for each action in the environment.

The action set, i.e., $\mathcal{A}$, of the EV in the grid world denotes the way how the EV can moves to change its state. In our scene, the directions that the EV is allowed to use are included in the following set:

$$\mathcal{A} = \{South, North, East, West, South-East, \\ South-West, North-East, and\ North-West\} \tag{12}$$

The energy consumption cost of EV movement towards CS, i.e., $\vartheta_{ev,j}$, is mainly dependent on the distance between the locations of EV and CS. To minimize this cost, we need to reduce the covered distance that the EV

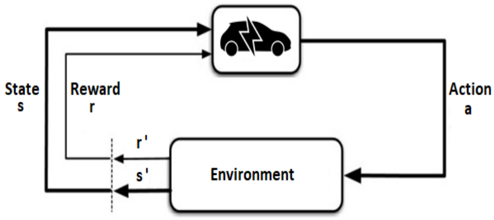

Fig. 2. Example of a single movement of EV

needs to travel to reach the CS. To achieve this, the EV earns punishment (penalty) for every hop in the grid. The rewards associated with CSs depends on the charging rates at these CSs. Accordingly, CS with higher charging rate, the lower its reward. The reward increases as the charging rate decreases. Fig. 2 shows an Illustrative example of a single movement of EV in the environment.

2) Q-Learning Algorithm: Two input parameters are required for Q-learning algorithm: the state in which the agent is located, and an action that can be taken at the current state. Therefore, the Q-learning algorithm has a function that calculates the quality of these two parameters, as shown below:

$$Q : S \times R \rightarrow \mathbb{R} \tag{13}$$

Before an agent starts its learning process, Q is initialized to an arbitrary fixed value (zero in our approach). Then at each step the agent chooses an action a_t, earns a reward r_t, moves to a new state s_t, then Q is updated. In this work, the objective function is minimizing the total energy cost of an EV which can be calculated using the optimal strategy by recursively updating action-value function (Q). The value of this function is determined by the TD learning algorithm technique and the Bellman equation, as shown bellow:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \; max_{a'} \; Q(s',a') - Q(s,a)) \tag{14}$$

where $Q(s,a)$ represents the value of the state-action function, i.e, Q-value function (s,a), α and γ denotes the learning rate and discount factor between 0 and 1, and r is the immediate reward value received as the result of taking action a in state s. In the environment, there are many different Q-value functions according to the different actions and policies that can be used in the learning process. The optimal Q-value function is the value which yields maximum Q-value compared to all other Q-values that have been acquired during the learning process. So, mathematically the optimal Q-value function, i.e, state-value function, can be expressed as:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{15}$$

where $Q^*(s,a)$ denotes the optimal value-action function.

The purpose of the EV charging navigation process, is to find the optimal policy π^* over all feasible policies that EV can select during the learning process, which minimizes the cost or maximizes the reward. Therefore, Once we have $Q^*(s,a)$, EV can act optimally based on the optimal strategy as shown in the following greedy strategy:

$$\pi^*(s) = \operatorname*{argmax}_{a} Q^*(s,a) \tag{16}$$

where $\pi^*(s)$ is the optimal policy that EV can act in the environment, which achieves the optimal value-action function.

Q-learning algorithm uses ϵ-greedy policy for the action selection step, which is also called behavior policy, to ensure a high level of balance between exploration and exploitation, i.e, exploration-exploitation trade-off, as well as to improve the learning level of the agent during the direct interaction with the environment. A simple strategy that has been proposed to deal with this problem is the ϵ-greedy (with $0 \leq \epsilon < 1$), with greater corresponding to greater probability of exploration. the value of ϵ has a significant impact on the performance and complexity of the Q-learning algorithm. Details can be seen in Algorithm 1 as follows:

Algorithm 1 Training Process of Q-Learning algorithm

Input: $\alpha, \gamma, \epsilon, Q$ (terminal-state), $s \in S, a \in \Lambda, N, M$
Output : Optimal $Q^*(s,a)$, and $\pi^*(s)$ for EV

Initialization:
```
Initialize Q(s,a) arbitrary
Initialize fixed CS locations
K=maximum number of episodes
Initialize random s for the EV
for eps = 1 to K do
    Select a ∈ Λ for EV using ϵ-greedy policy
    Execute the action a
    Receive immediate reward r
    Observe the new state s'
    Select a' in s' for the EV using Eq. (14)
    Update Q(s,a) value in Q-table
    s ← s'
    if (s is terminal or ϑ_{ev,j} ≥ Φ) then
        Start new episode
    else
        Select a ∈ Λ
    end if
end for
Return Q*−values, and π* for the EV
```

IV. Experiments

In this section, the evaluation of the proposed approach in different study cases are carried out within the proposed environment to demonstrate the effectiveness and feasibility of the proposed approach. Moreover, a comparison between the baseline case and the greedy strategy is also presented in this paper. In Section IV-A, the details of experimental setup are presented. The training process and simulation results are discussed in Section IV-B.

A. Experimental Setup

The performance of the proposed approach is demonstrated within an area 25× 25 grid map. The agent (EV) earns -1 as penalty for each movement in the environment. The rewards that the EV earns when reaching CSs varies depending on the charging rate of each CS which mainly depends on the electricity rates as mentioned earlier. Barriers have been placed on some of the roads that EV takes in the directions leading to CSs, which in turn force the EV to search for another available roads. In this work, we assume that the rewards of CSs are two values 60 and 80, depending on the charging rate at CS. As mentioned earlier, The reward increases as the charging rate decreases. An inverse relationship between the charging rate and the given reward associated with each CS. Each episode terminates, if the EV reached to the CS or reached the threshold value of the travelling energy consumption (Φ). All the following experimental results

TABLE I. RL-EVAS parameters

Parameter	Value
Environment	25×25 grids
M	4
Penalty	-1
Rewards	60, 80
ζ_{ev}	62 kWh
φ	\$0.15, \$0.35
SoC	ζ_{ev} × 60% kWh
α	0.1
ϵ	0.1
γ	0.6
δ_{ev}	0.16 kWh/km
Φ	1.6 kWh

have been performed by Python 3.10 on Windows 10 Pro 64bits, V.20H2, Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz (8 CPUs), 1.80 GHz. The simulation parameters related to the proposed scheme are presented in Table I.

B. Results

The performance of RL-EVAS is evaluated with respect to two criteria:

- The maximum cumulative reward of the value function of the learned policy, reflecting the proposed objective function of this approach.
- The total energy cost of EV, considering the cost that resulting from the movement of EV towards CS, and the cost of charging the EV at CS.

To this end, comparisons between RL-EVAS and different case studies, and also between the RL-EVAS and greedy strategy will be conducted in the proposed scheme.

1) Case Studies: The following proposed case studies demonstrate the feasibility and effectiveness of the proposed approach.

- Case A: With reduced the number of actions. In this case, we assume that the number of actions that the EV needs to perform to interact with the environment at each state is just 4 as shown in Eq. (17), rather than 8 actions for RL-EVAS, as shown in Eq. (12). As shown in Fig. 1, the green and blue arrows represent the possible actions that the EV can perform in Case A. While the possible actions that the EV can select in RL-EVAS are the green, blue and red arrows which give the EV to perform 8 actions. Finally, the other parameters remain the same as the RL-EVAS.

$$\mathcal{A}_{CaseA} = \{South,\ North,\ East,\ West\} \quad (17)$$

- Case B: With the increase in the number of obstacles standing in the way of the EV towards CSs. In this case, we assume that the number of barriers is increased by 25%, 50%, and 75% compared to the RL-EVAS with the same number of episodes. While the other parameters remain the same as the RL-EVAS.

Fig. 3 and Table II show the comparison results between RL-EVAS and Case A. It can be seen that the distance, travelling energy consumption, and travel time are less in RL-EVAS compared to Case A, this leads to minimize the total amount of energy needed to charge the EV. As a result, the total energy cost is minimized as shown in Fig. 3, this is due to the assumption that the possible actions that EV can select in RL-EVAS is more compared

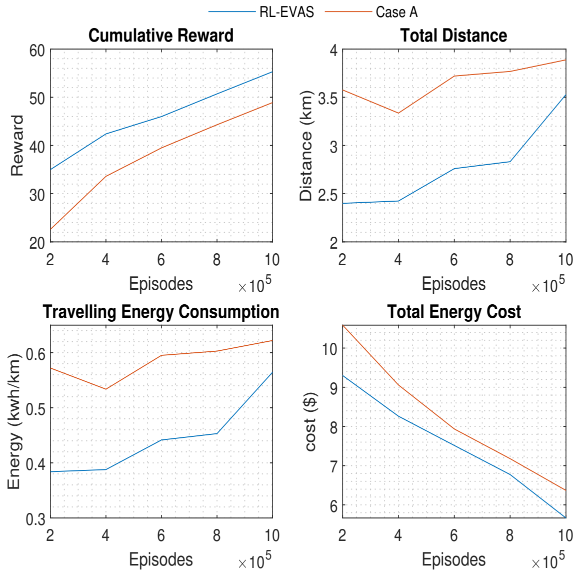

Fig. 3. Comparison between RL-EVAS and Case A

to Case A. Another observation in Fig. 3, is that the cumulative reward for RL-EVAS is more compared to Case A. The reason behind this is that the total number of timestep (the hops that EV needs to reach CS) is less compared to the Case A as shown in Table II. Each timestep is considered as an additional penalty for the EV. Accordingly, the cumulative penalty will finally be deducted from the reward that EV receives when it reaches CS. Based on the foregoing, it is noticeable that the cumulative reward is inversely related to the timestep.

Fig. 4 and Tables III - V, show the comparison results between RL-EVAS and Case B. In this case, we assume three different scenarios. In the first scenario, we assume that the number of obstacles is increased by 25%, then in the second scenario we assume that the number is increased by 50%, while in the last scenario, we assume that the number is increased by 75%. The reason behind this assumption is that street conditions are not fixed and can change for several reasons, including, but not limited to, the works that may occur in the study area as shown in Fig. 1.

It is easy to notice that that the total timestep, total distance, travelling energy consumption, and the total energy that is required to fully charge EV are less in RL-EVAS compared to the all scenarios in Case B, and it can be observed that all of these parameters are increased according to the proposed scenario in Case B as shown in Fig. 4. The reason for this is the assumption of increasing the percentage of streets that the EV cannot use in the study area to access CS. To overcome this challenge, the EV will try to find other possible routes to reach CS, which in turn increases the number of hops (timestep) that the EV must take to reach CS. Consequently, it increases the distance between the location of EV and the location of the CS at any charging decision point, the energy consumed by EV en route to the CS, and also the total energy that is required to fully charge EV. It is also noticeable that the timestep, distance, travelling energy consumption and total energy decrease when the number of episodes is

TABLE II. Comparison between RL-EVAS and Case A in terms of Timestep, Total Energy and Travel Time

Episodes	RL-EVAS			Case A		
	Timestep	Total Energy (kwh/km)	Travel Time (minute)	Timestep	Total Energy (kwh/km)	Travel Time (minute)
2×10^5	5	37.584	3.6	7.45	37.77216	5.364
4×10^5	5.05	37.58784	3.636	6.95	37.73376	5.004
6×10^5	5.75	37.6416	4.14	7.75	37.79519999	5.57985
8×10^5	5.9	37.65312	4.248	7.85	37.80288	5.652
10×10^5	7.35	37.76448	5.292	8.1	37.82208	5.831985
Average	**5.81**	**37.646208**	**4.9516**	**7.622**	**37.785215998**	**5.486367**

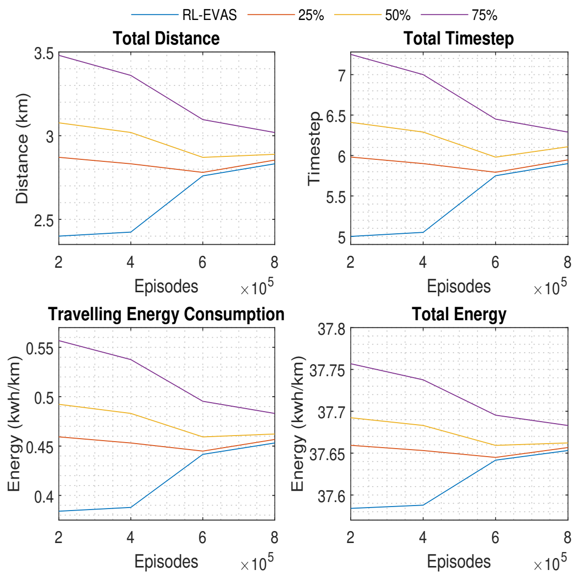

Fig. 4. Comparison between RL-EVAS and Case B

TABLE III. Comparison between RL-EVAS and Case B when the number of obstacles is increased by 25%

Episodes	RL-EVAS		Case B	
	Cumulative Reward	Total Energy Cost ($)	Cumulative Reward	Total Energy Cost ($)
2×10^5	35	9.3	34.27	9.8864
4×10^5	42.4	8.2613759	41.8	8.3891
6×10^5	46	7.515455	43.2	7.67741
8×10^5	50.7	6.772416	48.9	6.78912
Average	**43.525**	**7.9623117**	**42.0425**	**8.1855075**

higher as shown in Fig. 4. This is due to the fact that the performance of the EV in the study area improves with the increase in the number of episodes, since the EV will have a higher chance of finding the optimal CS at any charging decision point, even though the number of available streets decreases.

As shown in Tables III - V, the cumulative reward that has been achieved in RL-EVAS is higher compared to all scenarios of Case B, which means that the EV in RL-EVAS has selected CSs with higher reward rather than selecting CSs with lower rewards as the EV performed in Case B. As mentioned in Section III-B, the charging rate at CS decreases as the given reward increases. This is the reason why the total energy cost for charging EV at selected CSs in RL-EVAS is less compared to all scenarios in Case B as shown in the above-mentioned tables.

Tables VI, shows the comparison between the RL-EVAS and Case B, in terms of the travel time of EV to reach CS.

TABLE IV. Comparison between RL-EVAS and Case B when the number of obstacles is increased by 50%

Episodes	RL-EVAS		Case B	
	Cumulative Reward	Total Energy Cost ($)	Cumulative Reward	Total Energy Cost ($)
2×10^5	35	9.3	33.67	10.0879
4×10^5	42.4	8.2613759	38.9	8.4662
6×10^5	46	7.515455	44	7.62554
8×10^5	50.7	6.772416	47.9	6.82147
Average	**43.525**	**7.9623117**	**41.1175**	**8.2502775**

TABLE V. Comparison between RL-EVAS and Case B when the number of obstacles is increased by 75%

Episodes	RL-EVAS		Case B	
	Cumulative Reward	Total Energy Cost ($)	Cumulative Reward	Total Energy Cost ($)
2×10^5	35	9.3	32.87	10.32453
4×10^5	42.4	8.2613759	37	8.7986
6×10^5	46	7.515455	42	7.70554
8×10^5	50.7	6.772416	46.3	6.92374
Average	**43.525**	**7.9623117**	**39.5425**	**8.4381025**

It is seen that the travel time of RL-EVAS is less compared to the all scenarios in Case B, with an improvement of up to 20%. The reason behind this is that the number of the streets that the EV cannot use to find CS with is increased compared to RL-EVAS, which means that the EV needs to travel longer distance, which in turn increases the travel time to reach CS.

TABLE VI. Comparison between RL-EVAS and Case B in terms of the travel time (minute)

Episodes	RL-EVAS	Case B		
		25%	50%	75%
2×10^5	3.6	4.30618	4.6152	5.22
4×10^5	3.636	4.248	4.5288	5.04
6×10^5	4.14	4.17096	4.3056	4.644
8×10^5	4.248	4.28134	4.33318	4.52837
Average	**3.906**	**4.25162**	**4.4457**	**4.858075**

TABLE VII. Comparison between RL-EVAS and the Greedy Strategy in terms of Total Energy and Total Energy Cost

Episodes	RL-EVAS		Greedy Strategy	
	Total Energy kwh/km	Total Energy Cost ($)	Total Energy kwh/km	Total Energy Cost ($)
2×10^5	37.584	9.3	37.56096	10.51706
4×10^5	37.58784	8.26137	37.55788	9.78
6×10^5	37.6416	7.51545	37.53484	10.6891
8×10^5	37.65312	6.77241	37.49184	8.6871
10×10^5	37.76448	5.66467	37.47801	9.0247
Average	**37.64621**	**7.50278**	**37.52471**	**9.73959**

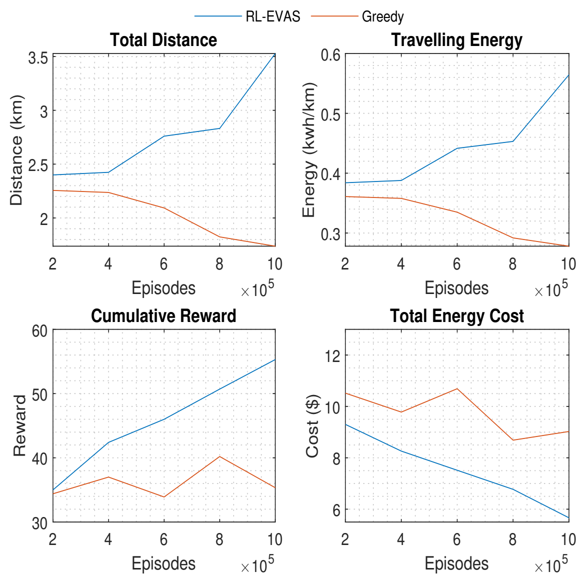

Fig. 5. Comparison between RL-EVAS and Greedy strategy

2) A Comparison between RL-EVAS and Greedy strategy: In this work, in addition to the proposed case studies that have been assumed before, we also compare RL-EVAS with greedy strategy to demonstrate our proposed approach, the comparison between the two strategies is done based on the total distance, travelling energy consumption, total energy that is required to fully charge EV, cumulative reward, and total energy cost. As mentioned before, the performance of RL-EVAS is evaluated with the respect into two criteria: the cumulative reward, and total energy cost. Fig. 5 and Table VII show the results of the comparison between RL-EVAS and the greedy approach.

Although the greedy algorithm achieved comparable results in terms of total distance, travelling energy consumption, total energy that is required to fully charge EV as shown in Fig. 5 and Table VII. However, RL-EVAS was able to achieve the desired results, in terms of maximizing the cumulative reward and minimizing the total energy cost cost. The reason for this is the greedy algorithm selected the CS based only on the distance, which in turn reduced the energy consumption to reach CS, and the total energy required to fully charge EV. However, the greedy algorithm did not take into account the variance in the rewards that have been associated with each individual CS, and also the charging rate at each CS, which in turn led to reduce the cumulative reward and increase the total energy cost. On the contrary, RL-EVAS has taken into account the two parameters, thus achieved the goal of the system, which is maximizing the cumulative reward and minimizing the total energy cost.

V. Conclusion

This paper proposed a RL-based assignment scheme for EVs to CSs. Several parameters were considered, including the distance between the locations of EV and CS, the EV travelling energy consumption, charging rate at each CS, and the total energy cost of EV. An EV's battery SoC was also considered as a constraint to limit the amount of energy consumption that an EV consumes to reach CS. Experimental results demonstrate that the proposed scheme can approximate the solution of finding the optimal policy in the sense of maximizing the expected value of the total reward, and minimizing the total energy cost of EV using Q-learning algorithm. Moreover, the results of the comparisons we obtained showed that the RL-EVAS outperformed all the proposed cases and greedy algorithm approach with an improvement of up to 20%. Our future works will focus on improving the scalability of the proposed scheme for practical applications, and using the Deep Q-Network (DQN) algorithm to solve this problem using multi-agents in the study area.

References

[1] M. Chowdhury, "A collaborative charge scheduling scheme for electric vehicles (ev) based on charging time and resource cost-awareness," *International Journal of Sensor Networks*, vol. 38, no. 2, pp. 122–131, 2022.

[2] Y. Zhang, S. Zhou, X. Rao, and Y. Zhou, "Ev-road-grid: Enabling optimal electric vehicle charging path considering wireless charging and dynamic energy consumption," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*. IEEE, 2021, pp. 1–5.

[3] M. Aljaidi, N. Aslam, and O. Kaiwartya, "Optimal placement and capacity of electric vehicle charging stations in urban areas: survey and open challenges," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE, 2019, pp. 238–243.

[4] M. R. Mozafar, M. H. Moradi, and M. H. Amini, "A simultaneous approach for optimal allocation of renewable energy sources and electric vehicle charging stations in smart grids based on improved GA-PSO algorithm," *Sustainable cities and society*, vol. 32, pp. 627–637, 2017.

[5] A. Chiş, J. Lundén, and V. Koivunen, "Reinforcement learning-based plug-in electric vehicle charging with forecasted price," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 3674–3684, 2016.

[6] K.-B. Lee, M. A Ahmed, D.-K. Kang, and Y.-C. Kim, "Deep reinforcement learning based optimal route and charging station selection," *Energies*, vol. 13, no. 23, p. 6255, 2020.

[7] L. Zhang, K. Gong, and M. Xu, "Congestion control in charging stations allocation with q-learning," *Sustainability*, vol. 11, no. 14, p. 3900, 2019.

[8] T. Qian, C. Shao, X. Wang, and M. Shahidehpour, "Deep reinforcement learning for ev charging navigation by coordinating smart grid and intelligent transportation system," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1714–1723, 2019.

[9] M. Aljaidi, N. Aslam, X. Chen, O. Kaiwartya, and Y. A. Al-Gumaei, "Energy-efficient EV charging station placement for E-mobility," in *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2020, pp. 3672–3678.

[10] M. Aljaidi, N. Aslam, X. Chen, O. Kaiwartya, and M. Khalid, "An energy efficient strategy for assignment of electric vehicles to charging stations in urban environments," in *2020 11th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2020, pp. 161–166.

[11] T. Ma and O. A. Mohammed, "Optimal charging of plug-in electric vehicles for a car-park infrastructure," *IEEE Transactions on Industry Applications*, vol. 50, no. 4, pp. 2323–2330, 2014.

[12] F. V. Cerna, M. Pourakbari-Kasmaei, R. A. Romero, and M. J. Rider, "Optimal delivery scheduling and charging of evs in the navigation of a city map," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4815–4827, 2017.

[13] M. Liu, Y. Wan, F. L. Lewis, and V. G. Lopez, "Adaptive optimal control for stochastic multiplayer differential games using on-policy and off-policy reinforcement learning," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 12, pp. 5522–5533, 2020.

[14] M. Dabbaghjamanesh, A. Moeini, and A. Kavousi-Fard, "Reinforcement learning-based load forecasting of electric vehicle charging station using q-learning technique," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4229–4237, 2020.