IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 6, 2021, accepted February 27, 2021, date of publication March 8, 2021, date of current version March 19, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3064354*

# Reinforcement Learning Based EV Charging Management Systems–A Review

**HEBA M. ABDULLAH, (Member, IEEE), ADEL GASTLI, (Senior Member, IEEE), AND LAZHAR BEN-BRAHIM, (Senior Member, IEEE)**
Electrical Engineering Department, College of Engineering, Qatar University, Doha, Qatar

Corresponding author: Heba M. Abdullah (ha1808917@qu.edu.qa)

This work was supported by the Qatar National Library.

**ABSTRACT** To mitigate global warming and energy shortage, integration of renewable energy generation sources, energy storage systems, and plug-in electric vehicles (PEVs) have been introduced in recent years. The application of electric vehicles (EV) in the smart grid has shown a significant option to reduce carbon emission. However, due to the limited battery capacity, managing the charging and discharging process of EV as a distributed power supply is a challenging task. Moreover, the unpredictable nature of renewable energy generation, uncertainties of plug-in electric vehicles associated parameters, energy prices, and the time-varying load create new challenges for the researchers and industries to maintain a stable operation of the power system. The EV battery charging management system plays a main role in coordinating the charging and discharging mechanism to efficiently realize a secure, efficient, and reliable power system. More recently, there has been an increasing interest in data-driven approaches in EV charging modeling. Consequently, researchers are looking to deploy model-free approaches for solving the EV charging management with uncertainties. Among many existing model-free approaches, Reinforcement Learning (RL) has been widely used for EV charging management. Unlike other machine learning approaches, the RL technique is based on maximizing the cumulative reward. This article reviews the existing literature related to the RL-based framework, objectives, and architecture for the charging coordination strategies of electric vehicles in the power systems. In addition, the review paper presents a detailed comparative analysis of the techniques used for achieving different charging coordination objectives while satisfying multiple constraints. This article also focuses on the application of RL in EV coordination for research and development of the cutting-edge optimized energy management system (EMS), which are applicable for EV charging.

**INDEX TERMS** Artificial intelligence, electric vehicles, machine learning, management, smart grids.

## I. INTRODUCTION

Thanks to the recent advancements in the battery industry, and the growing pressure from climate change and greenhouse gas emission reduction policies, the concept of plug-in electric vehicles (PEVs) has been prompted. The PEVs directly use electricity instead of fossil fuels, thus providing environmental and economic benefits [1]. The increasing numbers of PEVs will significantly affect the power systems' electric load profiles, creating dynamic and unpredicted demands for electricity that pose new challenges and constraints on the grid. Thus, developing high-efficiency energy management controllers for the coordination of EV charging systems is becoming a challenging task.

Numerous algorithms, such as rule-based and optimization methods, have been proposed in the literature for energy management systems [2]. Rule-based approaches are constructed by pre-defined policies, heuristics, or human expertise to calculate the energy distribution of multiple power sources. This method is highly reliable and robust but lacks the adaptability and flexibility to changing cycles. The optimization techniques formulate a mathematical model of an environment and calculate the optimal behavior systems. However, in the case of an EV charging management application, the system is highly dynamic, and optimization techniques cannot be used to evaluate the optimal behavior for EV charging management systems.

The literature is congested with extensive and detailed reviews analyzing the PEV charging impacts on the electric distribution system [3], [4], and [5]. Some researchers

The associate editor coordinating the review of this manuscript and approving it for publication was Mahdi Pourakbari Kasmaei.

highlighted the benefit of using vehicle-to-grid (V2G) technology, which is discharging some of the energy stored in the batteries of PEVs into the grid, to reduce the peak demand in the microgrid [6]. The literature review presented in [1] summarizes the impact of EV charging and presents the pros and cons of charging with V2G mode. The impact of V2G on electric power systems is also reviewed in [7]. Recent studies focus on analyzing V2G service impact on increasing the degradation process of the PEV battery [8].

The integration of PEVs poses several challenges to the operation and control of electric power systems. New PEV battery management systems need to be developed to overcome these challenges, to operate and control the charging mechanisms and energy flow in the grids with increasing and high penetration of PEVs. The coordinated charging problem is a relevant research topic that has been extensively studied in the literature, and various solutions have been [9], [10], [11]. For instance, survey papers [12] and [13] show that for the distribution networks to accommodate the increasing demand for PEV charging in the future, smart charging strategies must be put into operation. The studies related to PEV charging coordination are usually classified by their optimization objectives and charging approaches.

Researches in [14] emphasized the smart grid development to set forward PEV movement in a positive direction. The smart grid enhances EV coordination strategies by allowing communication and control. The smart grid is an electric power grid that uses two-way data exchange through communication technology with several objectives such as: to save energy, reduce cost, increase reliability and transparency, and enable new applications such as renewable distribution generation, electric-vesicles etc. [15]. Several surveys conducted on EV transportation with smart grid systems are analyzed and summarized in [16].

EVs interact with elements in an environment with high uncertainties such as; the status of the power grid, the demand of other EVs, EV arrival and departure times, battery state of charge, availability of chargers, price of electricity, charging preferences, etc. Furthermore, the integration of renewable energy sources causes high uncertainty because of their intermittency.

It was noticed that the literature pays less attention to the handling of uncertainties in the systems, which involve human behavior as in connecting/disconnecting loads and PEVs, or as unpredicted data such as the intermittency of renewable sources, etc. Since a smart grid is a largescale system, it is exposed to frequent changes such as grid voltages, EV arrival/departure times, frequency fluctuation, distribution system loading etc. In this case, large data requirement and lack of scalability makes centralized algorithms unfeasible to obtain global information and is relatively expensive to implement. For instance, [17] and [18] proposed a centralized strategy in real-time that runs a load flow to update data every 5 minutes. The network data is required for the control of commands sent to the chargers to avoid voltage violations and electrical equipment overloading. However, this method is unable to define individual EV driver's status or needs, such as energy requirement or arrival and departure times. Also, centralized coordination faces difficulties in responding to intermittent power sources such as solar or wind because of the frequent changes in the grid voltages. This creates a significant disadvantage in the centralized approach of EV energy management. A large number of EVs in the system need great computational power and large data-transfer rates.

The authors of [19] concluded that the correct understanding of future data can significantly improve the performance of online algorithms. A day-ahead (DA) charging scheduling method is proposed in [20]. While [21] suggests solving EV scheduling within the smart environment to minimize the generation and charging cost. A centralized approach by [22] minimizes power losses. However, none of these works addressed the uncertainty of driving demand or electricity prices, which are the smart grid's challenges [23]. Decentralized coordination can meet the driver's requirements while considering the grid and power quality. Decentralized control algorithms are deemed as a promising alternative for several reasons suggested in [24]. The previous studies based on the optimization approaches cannot perform well in case of high uncertainties and flexibility for modeling the behavior of EV.

Some researchers have also utilized the concept of Machine Learning (ML) to address the issue of EV energy management systems. In this case, real-time charging control is necessary, but because these parameters are highly variable, huge data sets, and uncertain prediction, the adaptation of ML for EV charging control has become crucial [25]. ML models can use both historical data as well as daily variables, such as weather and traffic data, to accurately capture the trends in charging behavior. The transportation sector acquires a large amount of data collected from multiple sources such as the Global Positioning System (GPS), closed-circuit television (CCTV), road sensors, event reports, etc. Handling such an amount of data for control and management is rather challenging if it is done through conventional methods. Therefore, recently ML approaches are becoming recommended in both the research field and industrial applications [26], [27].

More recently, there has been an increasing interest in data-driven approaches in EV charging infrastructure modeling and control. Consequently, researchers are looking to deploy predictive analysis methods for solving EV charging management problems with uncertainties. A comprehensive review in [25] presents the use of supervised and unsupervised ML as well as Deep Neural Networks (DNN) in the literature for charging behavior analysis and prediction. A challenge with ML predictive models is that they are trained with specific data sets, which can only be used for a particular, well-defined case. For instance, a predictive driving behavior model for a specific city is obtained by training a particular data set obtained for that location. Therefore, the same model cannot be reused for other geographical locations. On the other hand, algorithms that are based on Artificial Intelligence (AI) are powerful approaches with outstanding advantages in complicated decision-making problems [26] and can be applied to systems with ongoing changes.

According to [25], many studies showed the superiority of AI models over the probabilistic models in performance. Authors in [28] review the deep learning methods applied to the transportation sector, such as traffic incident processing, travel demand prediction, traffic signal control, and traffic flow prediction without covering the EV coordination problem. However, several existing ML-based techniques are highly dependent on the nature of the datasets and cannot perform well if introduced in dynamically changing environments. However, some new unsupervised ML techniques, using Deep Learning (DL) and Reinforcement Learning (RL) are more suited to such environments. RL utilizes trial-and-error in order to learn the best action for maximizing the reward. Deep Reinforcement Learning (DRL) integrates reinforcement learning with deep learning, providing models that interact with the environment while maximizing the long-term reward. Therefore, the deep reinforcement learning agent interacts with the real-time variations in the electricity prices and EV charging schedules, while providing online solutions. It also has different execution methods including the simplex and hybrid algorithms, in their applications in EVs.

The knowledge production within the research field of EV battery charging infrastructure management is accelerating at a tremendous speed while at the same time remaining fragmented and interdisciplinary. This makes it difficult to keep up with state-of-the-art technologies and to be at the cutting edge of research, as well as to assess the collective studies related to the PEV charging management area of research. Therefore, an updated literature review on this topic is more relevant than ever.

Although there exist several review papers with regards to EV charging, they do not focus on charging coordination from a data-driven approach, specifically with the application of learning-based coordination. For instance, the review in [29] classifies the literature in EV charging to infrastructure planning, charging operations, public policy, and business models. Various technologies and standards for EV charging is reviewed in [30] as a case study from Germany. The authors of [31] reviewed the models and algorithms related to the location problem of charging stations. The studies related to EV charging scheduling, forecasting, and clustering are summarized in [11]. In contrast, the survey work presented in this article is solely focusing on the RL concepts for EV charging coordination. This extensive review addresses the solutions, opportunities, prospects, and algorithms used in solving the EV charging management problem.

The key contributions of the paper are as follows:

- Classify different types of EV charging strategies and discuss the limitation within the environment of high uncertainties.
- Address the application of the RL-based approach in EV coordination and identify algorithms and objectives.
- Provide context and guidance to researchers seeking to apply RL-based approaches to solve problems of the charging schedules of the EVs.
- Discuss the limitations of the existing studies and identify future research directions.

## II. BACKGROUND

### A. EV CHARGING CONTROL STRATEGIES

This section presents the charging strategies classifications found in the literature. Coordination strategies are divided into two categories: centralized and decentralized, which depend on the kind of information available and communicated between the EV and power network [9]. In the centralized strategy, a central unit collects information about the grid, load, and power generation to decide on the charging or discharging of the EV fleet. The decentralized strategy is a distributed charging strategy where each EV driver decides on when to charge the battery, assuming there is no central controller.

Three state-of-the-art controlled charging schemes were introduced in [10]. Indirectly controlled charging, smart charging, and bidirectional charging. Controlled charging means that the EV charging process is coordinated according to the controlling parameters.

As the name suggests, indirectly controlled charging does not affect the charging parameters, like charging duration, in a direct way. However, it may control other parameters that can affect the charging parameters. For instance, the energy cost can be managed as part of the charging mechanism that will influence users to charge at required times throughout the day. The charging rates are defined based on the power system to avoid congestion, voltage violations, and overloading of electrical network equipment. In return, the charging parameters are controlled indirectly.

Smart charging coordinates the charging operation through managing sets of parameters to reach certain goals like minimizing power loss, reducing charging costs, reducing the burden on the grid, maximizing the operator's profit, etc.

Bidirectional charging is a mechanism that permits EVs to act as storage systems by allowing EVs to inject energy into the smart grid through a discharging process. This coordination scheme is deployed with an objective to stabilize the grid that has intermittent renewable energy sources.

In [11], the authors classify EV charging control schemes into three types of strategies. They are scheduling, clustering, and forecasting strategies. The paper explains the benefits and risks of each strategy and focuses on the EV scheduling strategy, which uses optimization objectives such as economic saving, peak reduction, network utilization, and market participation. The authors concluded that centralized charging faces a challenge due to high computation requirements and lack of scalability. Most of the researchers focus on decentralized strategies and hybrid centralized-decentralized management to reduce system costs. The clustering studies have shown adverse outcomes and led to difficulties with the assessment process and suggested using statistical measures to overcome this problem. The forecasting of EV demand is unpredictable, which is challenging in large-scale EV demand. The probabilistic methods for EV scheduling are compared with supervised/unsupervised AI methods. Finally, the review in [11] compares the strategies with summary, data used, region and software used.

Other studies focus only on centralized and decentralized control [31], [32]. Centralized coordination is impractical due to poor scalability and its requirement to know user behavior, while decentralized strategies are scalable but suffer from uncertainty issues [32]. On the other side, decentralized control requires accurate prediction and forecasting of consumer reaction, which is not possible in the case of a dynamic EV environment.

Combining EVs and intermittent renewable generation requires more flexibility because of the high fluctuation in renewable generation based on the weather conditions. A dedicated chapter in [33] discusses three main control methods which investigate the role of EVs in the modern electricity network project Intelligent Zero Emission Urban System (IZEUS) in Germany. There are three main control methods for integrating EVs with renewable energy generation: direct, indirect, and distributed autonomous control methods. For direct control, the scheduling operation is done by a third party, such as the system service provider or utility. For perfect control, information about the battery and consumer preferences are required. The high implementation of renewable sources in the power network caused an increase in the share of distributed generation. In return combining flexible loads with high penetration of distributed generation causes a high degree of uncertainty compared with the conventional situation. The direct control method used in this situation is the disconnecting of EVs from the grid zones that face breakdowns.

Indirect control is realized by applying incentives such as dynamic electric tariffs or time-of-use (TOU), which reflect on the power system status. In this case, the consumer has two choices: Reacting with the incentives or disallowing the operating schedule from controlling the EV locally. However, predicting the PEV's reaction to the applied incentives proposes a challenge in the indirect control method. This causes an increase in forecast errors and thus, points to the superiority of the direct control method over the indirect method in terms of reliability. Therefore, the advantage of direct control over indirect is better reliability, but at the same time, it requires more information and communication.

In autonomous control, there are no external control signals, and EVs only use internal sensors. The deviation in voltage and frequency can act as an indicator that shows if an increase in the generated power is required in the case of negative measurement and vice versa. Hence, the voltage and frequency measurements are related to the state of the system, making them suitable control signals. Some researchers proposed controlling PEVs using frequency [33]. Voltage measurements can also be used by EVs to observe the local situation in order to control the charging behavior optimally. This, in turn, leads to optimized integration of renewable energy resources and flattening the peaks resulting from simultaneous charging.

The heart of the EV economy is the control of the EV charging and discharging mechanism or what is called the control strategies, which are categorized into rule-based and approximation-based approaches [2]. Rule-based strategies are set without the requirement of prior knowledge of the system and solve the charging problem in real-time but do not necessarily reach the optimum solution. The other category of EV charging management strategies follows an optimization algorithm that solves an objective cost function with solutions such as off-line and prediction-based approaches, which require prior knowledge of the system and has a high computation cost. However, learning-based strategies can balance between having an online scheme and reaching an optimum solution. Machine learning approaches can be supervised, unsupervised, RL, Deep learning, and deep RL, see Fig. 1.

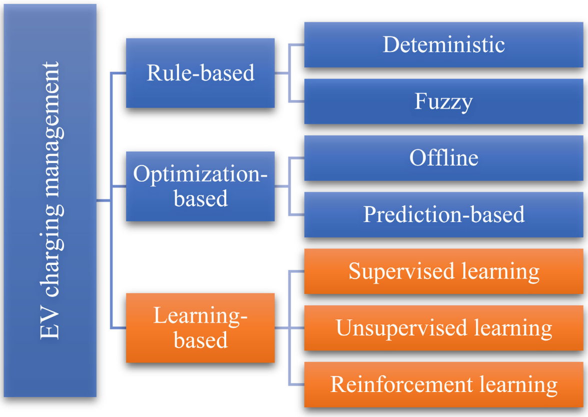


**FIGURE 1.** Energy management methods for electric vehicle charging coordination.

The most recent papers consider EV coordination with the EMS within a smart grid. The smart grid concept allows for energy exchange among PEVs, renewable sources, and energy storage systems. In this case, the EV coordination scheme becomes part of the EMS. A review of the concept of EMS and the different objectives and control architectures is presented in [34].

### B. MACHINE LEARNING

Machine learning (ML) is the program's ability to learn from experience through a training algorithm. The classification of ML approaches is in Fig. 2. ML methods used for obtaining the behavior of EV charging are reviewed in [25]. The paper explains in detail the ML approaches for predicting the charging behavior and does not discuss the coordination problem. Initially, ML programming starts with the learning process, which involves finding the trend and patterns in the training dataset and ends with the model that can make accurate predictions. These ML algorithms are categorized as supervised unsupervised. Another approach to teaching

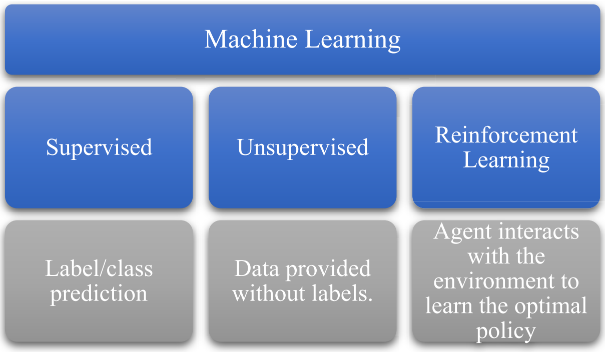

Machine Learning
Supervised
Unsupervised
Reinforcement Learning
Label/class prediction
Data provided without labels.
Agent interacts with the environment to learn the optimal policy

**FIGURE 2.** Machine learning approaches and definitions.

a machine by rewards when it achieves the expected job regardless of having a model is called RL.

There are two types of machine learning approaches supervised learning (SL) and unsupervised learning (UL) [35]. A continuous response variable is called a regression problem, while a categorical response variable is called a classification problem. Energy is therefore taken as the continuous variable, while night/day charging is a categorical variable. The learning process of SL or backpropagation approaches is defined in the literature as forward neural networks (FNNs) and recurrent neural networks (RNNs). The system parameters need to be known to supervise the training of these types of neural networks (NNs). In SL, the training data sets contain input variables and corresponding response variables called target variables. The SL is grouped by different learning styles such as linear regression, decision trees, random forest, support vector machine, k-nearest neighbor. Unlike SL, UL are statistical models that use training data sets of input variables only without target variables. This type of learning is deployed for clustering EV consumer behavior patterns. The types of UL are k-means clustering, Gaussian mixture model and Kernel density estimator.

A generalized type of ML is the RL, an approach where an agent learns from the environment from previous actions and does not need a model of the environment. After periodically making action decisions, the algorithm learns and observes the returned rewards, which is trial and error. The process is repeated by adjusting the policy of selecting certain actions to find the optimal strategy, yielding to the highest sum of rewards.

Another type of learning is called deep learning, and it is the composition of a large number of learned functions. It specifically uses hierarchy concepts that are defined by simpler and abstract representations. Multilayer Perceptron (MLP), Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN) are examples of DL.

In the concept of energy management and control, the SL approach assumes that events are independent of previous events, while this cannot apply in real-world problems that have dependable events. For instance, the optimal strategy for coordination of electric vehicles' charging is a sequential decision-making problem. This is because the parameters involved in EV charging coordination depend on previous events in time. In [25], RL techniques was used to provide solutions for EV problems concerning complex scheduling, leading to the application of RL techniques in [26] concerning smart grid. The smart grid application was mainly about scheduling the charging in dynamic urban environments. Therefore, the literature is surveyed to find the RL approach-based used for EV charging.

### C. REINFORCEMENT LEARNING

The learning agent in RL techniques observes the input state parameters obtained from the environment and chooses certain actions in order to maximize the reward. The learning agent in the EV charging problem is the charging management controller, which uses trial and error to decide on the best actions to execute in order to achieve the largest cumulative reward. Actions can affect the current and future rewards, so the cumulative reward is considered in RL algorithms. Therefore, the RL algorithm aims to achieve a certain goal which is reflected in the reward, that the learning agent tries to maximize through choosing the best actions depending on the observed states [36].

This article presents a comprehensive survey of the recent studies in the learning-based approaches in EV charging management. The application of RL algorithms in energy management can be generally classified into two categories: Simplex and hybrid algorithms. Simplex algorithms are algorithms in which the policy of energy management is derived using only one algorithm, such as Q-learning, Sarsa and Dyna algorithms [37]. On the other hand, hybrid algorithms integrate other types of algorithms with RL algorithms to derive the policy. These algorithms may include predictive algorithms, deep learning and model predictive control (MPC) [38]. The integration of additional algorithms leads to a more realistic and efficient approach to control EVs in real-time.

#### 1) MATHEMATICAL REPRESENTATION

Reinforcement learning learns by trial and error to reach an optimal policy [37]. The application of RL is effective for non-stationary environments that change over time. This is applied to the RL algorithm as a constant weight $\alpha$ to make the recent reward heavier than past ones. RL can be applied to a sequential decision task such as the discounted episodic Markov Decision Process (MDP).

The reinforcement learning cycle is illustrated in Fig.3, which shows the basic elements of an RL system. In an MDP problem $(S, A, \gamma, P, r)$, where $S$ is a set of all possible states, $A$ is the set of all possible actions at states$s_t$, and the reward R is discounted and cumulative reward using a discounted factor $\gamma \in (0, 1]$. In the RL learning process, for every episode, the next state $s_{t+1}$ is based on the transition probability $P(s_{t+1}|s_t, a_t)$ and the reward. Policy is the function that maps the new state to the selected possible action $a_t$. Policy $\pi_t(a|s)$ is the probability that $a_t = a$ and $s_t = s$.

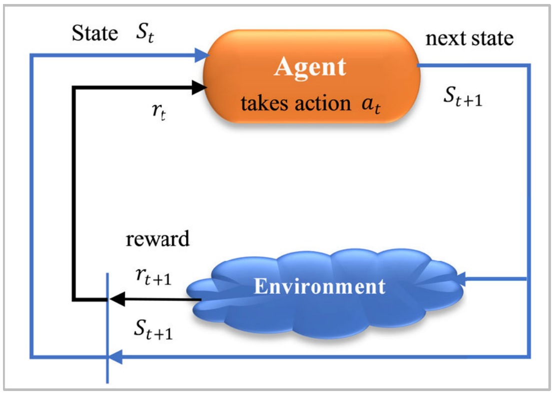


**FIGURE 3.** Illustration of the reinforcement learning cycle.

At each time step $t$, the agent receives the state $s_t \in S$ from the environment. Then, the action $a_t \in A$ of state $s_t$ is chosen by the agent using a policy $\pi_t(a|s)$ to maximize the reward $r_{t+1} \in R$. The environment, in its turn, sends an updated $r_{t+1}$ reward as a reaction to the agent's chosen actions at every time step. The process is repeated until the termination state or time step has reached its limit [37].

A complete probability distribution of the environment's respond at time $t+1$ to an action taken at time $t$ depends on previous actions and rewards.

$$P_r\{s_{t+1} = s', r_{t+1} = r|s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0|,\} \quad (1)$$

MDP has a clear basis of choosing the action and new state $s_{t+1} = s$ at $t+1$ considering the state $s_t$and action $a_t$ at $t$ only without the need of previous states and actions, where $s'$ is the next new state value.

$$P_r\,\{s_{t+1} = s, r_{t+1} = r|\ s_t, a_t|\ ,\} \quad (2)$$

When the agent follows a certain policy, the expected value of the state $V^\pi(s)$ for an MDP problem is called the state-value function for the policy $\pi$. The expectation when using policy $\pi$ is denoted as $E_\pi$, and $i$ is the iteration number before end of an episode [37].

$$V^\pi(s) = E_\pi\,\{R_t|s_t = s\} = E_\pi\left\{\sum\nolimits_{i=0}^{\infty} \gamma^i r_{t+i+1}|s_t = s\right\} \quad (3)$$

The value of the action $a$ under the same given policy followed by the environment in state $s$ is called the action-value function for policy $\pi$ denoted by $Q^\pi(s, a)$.

$$Q^\pi\,(s, a) = E_\pi\left\{\sum\nolimits_{i=0}^{\infty} \gamma^i r_{t+i+1}|s_t = s, a_t = a\right\} \quad (4)$$

The objective RL algorithms is to maximize the long-term reward and this is done by policy iteration. The concept of policy iteration is the bases of RL algorithms and consists of three steps: 1) Policy initialization, 2) Policy evaluation and 3) Policy improvement [36].

RL algorithms can be classified into two groups; value-based and policy-based methods. For value-based, the agent learns policy by updating the value function, and for policy-based, the policy is learned straight forward.

Monte-Carlo method (MC) and Temporal-Difference method (TD) are RL methods. The TD method has a faster convergence than MC because it does not wait to make an update at the end of the episode, but rather it makes an update within an episode [37].

TD method is divided into two categories:1) on-policy TD control (Sarsa, Actor critic) 2) off-policy TD control (Q-learning). The state-value functions are extended to the action-value function, decomposed to action and state, using the Bellman equation. On-policy means that the selecting policy and learning policy are the same. Where else in the off-policy method, they are the different. The mathematical representation for $V^\pi(s)$ and $Q^\pi(s, a)$ of the above RL methods are summarized in Table 1.

**TABLE 1.** Sate and action values of model free RL algorithms [36].

| RL learning method | $V^\pi$ / $Q^\pi(s, a)$ for current step $i$ in episode |
|---|---|
| Simple Monte Carlo | $V(s_t) \leftarrow V(s_t) + \alpha\big(R_t - V(s_t)\big)$ (5) |
| Simplest TD | $V(s_t) \leftarrow V(s_t) + \alpha\left(r_{t+1} + \gamma V^{i-1}(s_{t+1}) - V^{i-1}(s_t)\right)$ (6) |
| Sarsa | $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\big(r + \gamma\, Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\big)$ (7) |
| Q-learning | $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(r + \gamma\, max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right)$ (8) |

Q-learning is a widely used off-policy algorithm and it is a value-based method. This method obtains all possible actions of the next step and chooses the maximum Q, where else Sarsa uses the next values' estimation. These are RL basic learning algorithms referred to as one-step return. Online incremental method is called eligibility trace such as Kernel-based RL (KBRL) [37]. Deep neural networks can approximate RL components in deep RL method such as Deep Q-Networks (DQN) and Deep Deterministic Policy Gradient (DDPG). RL algorithms are improved by value function approximators, policy gradient and actor-critic to accelerate learning [37].

## III. LITERATURE ANALYSIS APPROACH

### A. CRITERIA FOR INCLUDING STUDIES

The method used to conduct this critical review paper is proposed in [39] to present a detailed survey and analysis of the peer-reviewed literature related to our topic, EV charging management, and discuss the theories presented in the surveyed papers along with their hypotheses, methods, and results if they exist. The advantages and drawbacks of certain methods are discussed along with their practicality in solving the EV management problem. The critical review follows a systematic review, as illustrated in Fig. 7.

First, the criteria, of included studies, are defined according to the research topic: EVs charging coordination based on unsupervised machine learning methods specifically

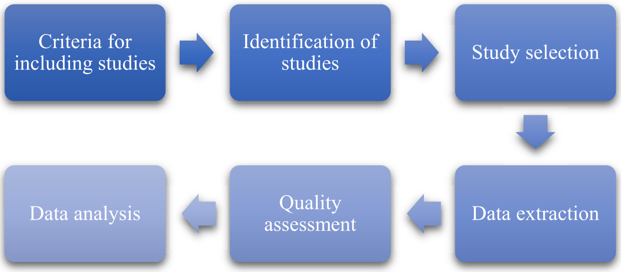


**FIGURE 4.** Process followed in systematic review.

the RL-based approaches. Firstly, we define and consider the keywords relevant to the topic, see Table 2. A research term is a combination of all main terms: electric vehicle, charging strategy, coordination method and objectives, with a selective term either from Group 1 or Group 2. Table 2 presents the keywords that were used to find papers in peer-reviewed journals related to the topic in online databases.

**TABLE 2.** List of search terms.

| | | Group 1 | Group 2 |
|---|---|---|---|
| | Electric Vehicle | "Electric vehicle" | "transportation"<br>OR "PEV"<br>OR "HPEV" |
| AND | Charging strategy | "charging"<br>OR "centralized"<br>OR "online"<br>OR "schedule"<br>OR "smart"<br>OR "distributed"<br>OR "decentralized" | "discharging"<br>OR "V2G"<br>OR "renewable"<br>OR "storage"<br>OR "energy"<br>OR "demand response" |
| AND | Coordination | "coordination"<br>OR "Q-learning"<br>OR "Reinforcement"<br>OR "RL"<br>OR "unsupervised"<br>OR "DRL"<br>OR "AI"<br>OR "ML" | "uncertainty"<br>OR "data"<br>OR "predictive"<br>OR "model"<br>OR "future" |
| AND | Objective | "electricity price"<br>OR "losses"<br>OR "battery"<br>OR "cost"<br>OR "load"<br>OR "satisfaction"<br>OR "peak"<br>OR "SoC" | "grid"<br>OR "smart grid"<br>OR "operator"<br>OR "transformer"<br>OR "aggregator" |

Keywords were selected according to characteristic features of RL and EV charging problems. The two groups of keywords were selected based on the inputs, constraints, objective, coordination scheme, and the algorithm of the RL-based framework used for the EV charging problem.

In the following two steps ($2^{nd}$ & $3^{rd}$), the electronic database considered in this search is defined as (Scopus, ISI Web of Knowledge, Google Scholar). The studies were identified based on the search results. In the next step, the studies resulted from the search in the online databases were examined by screening the abstract and selected for surveying if the results of these paper were considered beneficial and promising.

The fourth step in the survey is the data extraction, it includes classifying the information obtained in each research according to methods, objectives, algorithms, and contributions. Analysis of the studies selected was performed, and the advantages, disadvantages, contributions, further improvements, practicality, and several other aspects of the EV management problem were discussed and evaluated. Then the work's quality is assessed according to the published journal rank, the author's scientific contribution, paper number of citations, methods used in the research work, highlighted weakness, and the authors' contribution.

In the final step, we sort the studies and outcomes into different subtypes and classifications and compare work outcomes accordingly. Therefore, this survey paper provides a detailed discussion of the RL-based solutions used in the EV management problem and explains the drawbacks of traditional approaches to the problem while guiding scholars to improve the existing methods.

## IV. STUDIES REVIEWED

### A. PAPERS INCLUDED IN THIS STUDY

The papers included in this critical review include journal articles, conference papers, reports and PhD dissertations. A geographical distribution of the contributing countries of the corresponding authors' affiliations is plotted in Fig. 8. The most contribution comes from China, the United States, Canada, then Europe. At the same time, China has the World's largest EV market shares and global electric car sales. It is logical to say that counties with accelerating sales of EVs subsequently are contributing to the research field related to EV charging coordination with advanced artificial techniques based on the RL approach.

### B. EXISTING LITERATURE ON HANDLING CHARGING COORDINATION WITH UNCERTAINTIES

Modeling and simulation are useful for EV coordination applications, but they contain assumptions and uncertainties, which can cause accuracy limitations. Uncertain PEV load demand is considered one of the challenges in EV charging scheduling approaches.

As addressed in [40], finding the global optimal with a large population and random EV arrivals is not practical without the information of future loads, arrival times and charging periods. The study suggests a scalable distributed charging mechanism that finds a local optimal from individual EVs over a current set in a local group. The study in [40] is a simulation-based evaluation that proposes a local optimal scheduling scheme instead of a globally optimal scheduling scheme. The information required for system operation such as feeder loading, network congestion, EV demand are often unavailable to make informed decisions. Therefore, a stochastic approach is used instead of deterministic models to address uncertain PEV load demand [41].

Another way of handling the charging optimization problem with uncertainties from driver demands is by using

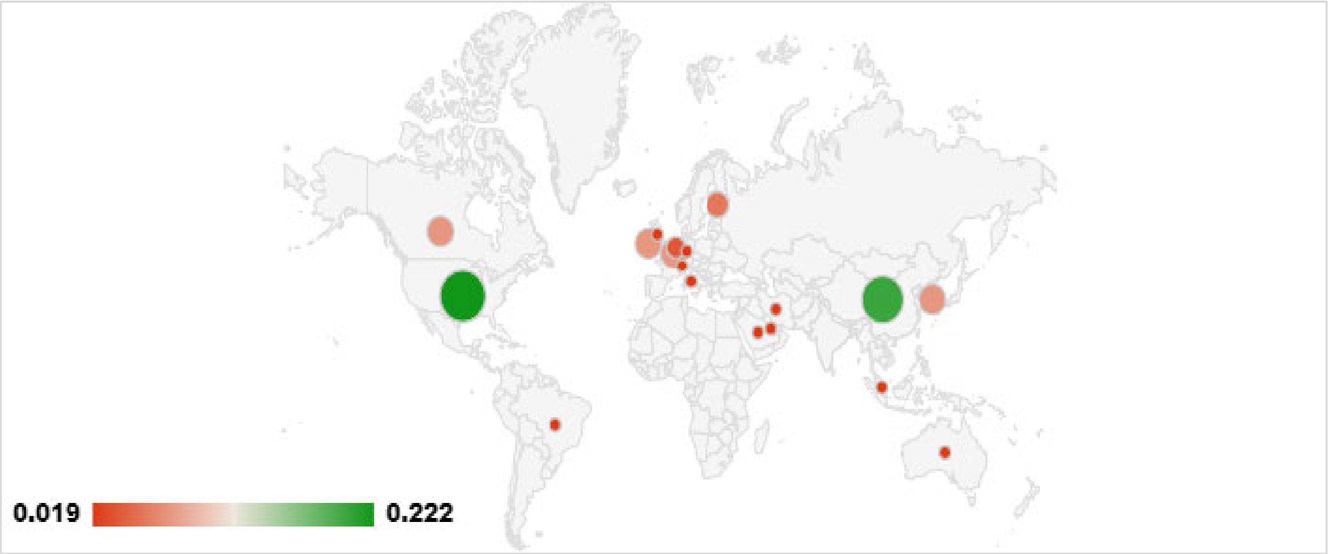


**FIGURE 5.** Geographical distribution of the affiliating countries for the contributing papers.

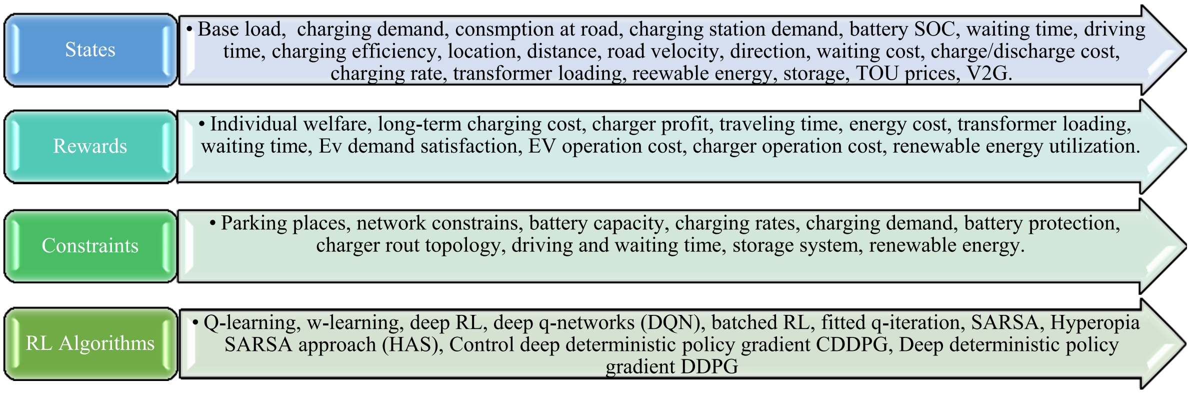

States
• Base load, charging demand, consmption at road, charging station demand, battery SOC, waiting time, driving time, charging efficiency, location, distance, road velocity, direction, waiting cost, charge/discharge cost, charging rate, transformer loading, reewable energy, storage, TOU prices, V2G.
Rewards
• Individual welfare, long-term charging cost, charger profit, traveling time, energy cost, transformer loading, waiting time, Ev demand satisfaction, EV operation cost, charger operation cost, renewable energy utilization.
Constraints
• Parking places, network constrains, battery capacity, charging rates, charging demand, battery protection, charger rout topology, driving and waiting time, storage system, renewable energy.
RL Algorithms
• Q-learning, w-learning, deep RL, deep q-networks (DQN), batched RL, fitted q-iteration, SARSA, Hyperopia SARSA approach (HAS), Control deep deterministic policy gradient CDDPG, Deep deterministic policy gradient DDPG

**FIGURE 6.** Summary of the parameters involved in the EVs charging coordination problems solved using RL approaches.

dynamic charging, model predictive control (MPC), and driving control solution as simulated in [42]. Deploying MPC-based feedback control is used to convert the problem from optimization to solving for least-squares. An MPC technique optimizes the plant's future behavior by calculating future adjustments as a sequence of controlled variables, and this process is iterated through the following time intervals; surveyed in [43]. The studies in [44] and [45] theoretically conclude that MPC is useful for handling uncertainties associated with distributed analysis.

Most common approaches have widely adopted MPC over dynamic charging and cost-driving schemes. [46] assumes the demand for the load is periodic in time and presents a plug-and-play approach based on MPC, aiming to bring the fluctuations in voltage to a minimum. Two studies [45] and [44] minimize load variation by predictive deferrable load control.

Not considering the load demand distribution, an online MPC approach is proposed in [19] instead of using generic convex optimization. The arrival times of EV are considered through statistical modeling. Dynamic programing with finite-horizon is used to model the problem of providing an online schedule. The model has continuous state-space and action-space. The future demand is unknown, but rather, the estimation of its statistic information is used where the charging cost is a function of the load demand. The charging coordination objective is to minimize charging cost and impact on the grid.

Dynamic programming approaches by [47] utilize iterative methods and can obtain optimal results. However, this method cannot predict under uncertainties. On the other hand, the day-ahead scheduling approach in [48] and real-time scheduling approach in [49] minimize uncertainties by robust optimization [50], [50] or stochastic optimization [51]. Real-time scheduling in [49], [49] is a model-based approach that highly depends on modeling the environment and the system parameters. However, deep learning (DL) in neural networks, reviewed by [35], uses an automated learning procedure that does not include human interference and is of hierarchical nature.

The existing techniques for handling the charging coordination with uncertainties cannot perform well with the flexible and uncertain characteristics of the EV model. Thus, a research gap exists in order to accurately model the charging coordination with uncertainties. Hence, the next section introduces the concept of RL to address the EV charging problem with different RL techniques in the literature.

### C. THE EV CHARGING MANAGEMENT PROBLEM WITH REINFORCEMENT LEARNING APPROACH

This section explains the recent techniques of the RL used in the EV charging management problem. The novelty of the survey is to explain the RL approach in EV charging management problems with respect to explaining the states, rewards, constraints, and algorithms used in the recent work.

The EMS's primary function for EV charging is to balance the power distribution of different power producers or consumers with an optimization goal. The EV coordination problem is part of the EMS used to control the charging parameters and achieve specific objectives such as operation objectives or financial satisfaction.

Operational performance objectives mentioned in the literature are, for example, load curve flattening, fairness in charging, minimizing voltage variation, avoiding transformer overloading, etc. The financial objectives are also considered as user satisfaction which is a type of objective that aims to minimize operating costs. Examples of financial objectives are charging cost minimization, power loss minimization, operator profit maximization, electrical network utilization maximization, etc. A summary of the inputs or states, objectives constraints, and RL algorithms found in the reviewed literature is illustrated in Fig. 9.

Solving the problem of managing the charging EVs is mostly an optimal control problem [52]. To operate the emerging grid optimally and efficiently, different methods were used in energy management systems or EMS. The strategy of solving energy management problems begins with gathering inputs and using them in an optimization framework to achieve economic and technical objectives under some constraints using some RL algorithms [53].

#### 1) INPUTS

To develop an RL solution, some parameters define the state of the environment, which the agent read as inputs to take actions to change the state of the environment toward the desired goal. After surveying several papers (see Table 3), many parameters in the environment have been used depending on the desired objectives of those papers.

Some parameters have been used by many researchers as state features of the environment. Their recurrent usage may imply that these parameters significantly represent the state of the environment. Another implication is that they are the parameters most needed by the agent to achieve the desired objective set by the specific study. For instance, studies that are concerned with meeting the EVs' charging demands used the battery SOC and the EVs' demands as state parameters such as [55] and [57]. Likewise, studies that are concerned about the energy cost or charging prices, took the operation cost, charging and discharging costs, TOU, energy prices as state parameters such as [56], [61], and [104].

Research targetting the balancing of the load profile and flattening the peaks deploy V2G as a state parameter to monitor EVs' availability that can discharge into the system and act as energy sources. Some other studies focus on using renewable energy sources and their inflexibility in producing energy, monitored their energy production to achieve better managing of the distributed energy systems. Other studies maintained the battery SOC of the EVs above a certain threshold by monitoring the battery SOC to be able to produce the appropriate actions [60], [74], [91], [92], [98], [99], [100], [102]. Other parameters were used less frequently in studies as state parameters, as they were not related to the frequently desired objectives [65] and [76]. The pattern in all these approaches is that you observe the parameters that you need to monitor to achieve your goal, and the agent learns how to optimize its actions through the action-reward system.

The authors in papers [56]–[65], [67], [69], [71], [72], [74], [75], [77], [79]–[84], [86], [88], [90]–[92], [94], [95], and [97]–[99], used the charging or discharging cost as a state parameter. However, other operational costs, like waiting cost and traveling, were used mainly in two of the surveyed papers [66] and [92]. This may lead to concluding that they were not that crucial in deducing the best actions to achieve the objectives of the methods presented in these papers to control the energy distribution system costs. TOU, on the other hand, appeared in several published research work ([59], [68], [70], [73], [81], [86], [87], [91], [93], and [104]) as a state parameter, but not as frequent as the charging or discharging cost

Methods proposed in [86] and [87] used both charging or discharging, cost and TOU as state parameters, which may imply that TOU is a good parameter for observing costs. TOU was used with V2G as state parameters in [59], [68], [86], and [104]. Only [104] had the objective of balancing the power usage or flattening the peaks, the other three papers aimed for either controlling energy cost, or, like [59], maximizing the distribution system's profit. V2G, as a state parameter, was used in [59], [60], [63], [64], [67]–[69], [75], [83], [86], [97], [100], Like TOU, V2G was used by methods that aimed to control the energy cost or maximize the profit of the distribution system.

Another parameter that was used as a state parameter by many of the proposed methods in [54], [57], [58], [63], [67], [68], [71], [72], [74], [75], [77], [78], [80]–[82], [85], [86], [88]–[90], [93], [94], [96], [97], [99], and [100]–[102], is the EVs charging demand satisfaction. These proposed methods aimed at controlling the energy cost, reducing the prices for the customers, or even increasing the charger's profit. The relation between the demand and the prices is existent in any business, so, eventually, it exists in the

**TABLE 3.** Summary of the RL parameters in the EV coordination problems found in the studied literature.

| Reference | Algorithm | States | Rewards | Constraints | Agent |
|---|---|---|---|---|---|
| [54] | Q-Learning | Energy Consumption Levels | **Single objective:** Increase individual welfare | EV charging constraints | Single |
| [55] | Deep RL, deep Q-network DQN and DDPG | Battery SOC, EV location and direction, number of occupied parking spaces and chargers | **Single objective:** Reduce charging time and fares | Parking places, charging availability | Single |
| [56] | Batched RL, fitted Q-iteration | Battery SOC, Electricity Prices and Price Fluctuations | **Single objective:** Reduce energy prices | Battery capacity, charging rate | Single |
| [57] | Batched RL, fitted Q-iteration | Required Energy, Charging Parameters | **Single objective:** Energy cost | Power consumption parameters, unpredictability | Single |
| [58] | Hyperopia SARSA approach (HAS) | Charging Demand, Parking Time | **Single objective:** Maximize charging station profit | Charging rates | Single |
| [59] | Deep RL, deep Q-network DQN and DDPG | Arrival/departure time, weather and traffic information, charging station load | **Single objective:** Distribution system profit | EV physical constraints, network constraints, charging constraints | Single |
| [60] | Q-Learning | Battery SOC, price, remaining time till departure | **Single objective:** EV owner's profit | Battery and charging constraints | Single |
| [61] | Q-Learning | Time till departure, battery SOC, renewable energy, energy price, type of user | **Single objective:** Charging station profit | NA | Single |
| [62] | SARSA | Battery SOC, Energy price | **Single objective:** Minimize charging cost | EV energy consumption | Single |
| [63] | Deep RL, deep Q-network DQN | Electricity Price, Battery energy levels, and time till charging | **Multi-objective:** Minimize charging cost, satisfy charging demand | Battery capacity | Single |
| [64] | Safe Deep RL | Battery SOC, Energy prices | **Single objective:** Maximize EV's owner profit | Charging demand satisfaction | Single |
| [65] | Deep RL, deep Q-network DQN | Traffic information, waiting time, charging prices | **Multi-objective:** Minimize energy cost and waiting time | Battery SOC and consumption, waiting and driving time, charging cost | Single |
| [66] | Decentralized learning, with centralized decision making using RL | Battery SOC, EV location, time till EV is available | **Multi-objective:** Minimize electricity costs, waiting time, EV operation cost | Charging station capacity | Multi |
| [67] | Control deep deterministic policy gradient CDDPG | Battery SOC, Energy prices, Arrival/departure time | **Multi-objective:** Maximize the EV owner's profit, Satisfy EV charging demand | Arrival/departure time, Battery energy consumption | Single |
| [68] | Q-Learning | Energy prices, EV charging demands | **Multi-objective:** Maximize the EV owner's profit, Satisfy EV charging demand | Physical battery constraints | Multi |
| [69] | Deep RL, deep Q-network DQN | ESS states, electricity price, load | **Single objective:** Maximize EVCS profit | ESS constraints | Multi |
| [70] | MASCO | Battery SOC, EV location, Transformer load | **Multi-objective:** Satisfy EV charging demand, Reduce energy costs, Reduce transformer overloading | Transformer load, number of friendly agents | Multi |
| [71] | W-learning | Battery SOC, System load, Transformer load | **Multi-objective:** Satisfy EV charging demand, Minimize operation costs, Reduce transformer overloading | Transformer load | Multi |
| [72] | Distributed W-learning | Wind-energy generation, load information, battery SOC | **Multi-objective:** Satisfying EV charging demands, Maximize Renewable energy use | Renewable Energy constraints | Multi |
| [73] | Q-Learning | Battery SOC, EV location | **Multi-objective:** Reduce waiting time, Satisfy EV charging demand, Increase the revenue | NA | Multi |
| [74] | Q-Learning | Battery SOC, Time till departure | **Multi-objective:** Satisfy EV charging demand, minimize charging cost | Charging constraint | Multi |

**TABLE 3.** ***(Continued.)*** **Summary of the RL parameters in the EV coordination problems found in the studied literature.**

| | | | | | |
|---|---|---|---|---|---|
| [75] | Q-learning | Battery SOC, Energy price, charging time | **Single objective:** Minimize charging cost | NA | Single |
| [76] | Batched RL | Temperature, Energy Consumption, Renewable energy Production, Battery SOC | **Single objective:** Reducing peak power consumption | Infrastructure constraints | Multi |
| [77] | Q-Learning | Battery SOC, time till departure | **Multi-objective:** Minimize energy cost, Satisfy EV charging demand | Network constraints, EV energy requirements | Multi |
| [78] | Distributed W-learning | System Demands | **Multi-objective:** Satisfy EV charging demand, Transformer load balancing | NA | Multi |
| [79] | Batched RL, fitted Q-iteration | Battery SOC, charging cost, | **Multi-objective:** Minimize charging costs, avoid battery depletion | Driving pattern constraints | Single |
| [80] | Q-Learning | Energy price, temperature, weather | **Multi-objective:** Reduce energy costs, customer satisfaction | NA | Single |
| [81] | Q-Learning | EV charging demand, arrival/departure time | **Single objective:** Reduce energy costs | EV owner's constraints | Single |
| [82] | Q-learning | Baseload, weather information | **Multi-objective:** Balance load profile, Satisfy EV charging demand | NA | Multi |
| [83] | Q-Learning | Battery SOC, EV mobility, charging demands | **Single objective:** Satisfy Micro grid demand, reduce energy prices | Battery SOC | Single |
| [84] | Deep RL, deep Q-network DQN and LSTM | Battery SOC, Energy prices | **Multi-objective:** Reduce energy cost, Satisfy EV charging demand | Battery constraints, arrival/departure time | Single |
| [85] | Batched RL, fitted Q-iteration | EV requested energy, arrival time, time till departure, charging rate | **Multi-objective:** Balance load profile, Satisfy EV charging demand | Parking places | Single |
| [86] | Q-Learning | EV charging demand, Battery SOC | **Single objective:** Reduce energy costs | ESS capacity | Single |
| [87] | Q-Learning | Energy consumption, ESS state of energy | **Single objective:** Reduce energy costs | ESS constraints | Single |
| [88] | Deep RL, deep Q-network DQN | Energy prices, charging demand, renewable energy generation | **Single objective:** Reduce energy costs | Physical grid constraints, transformer capacity | Multi |
| [89] | Deep Neural Network DNN | Energy price, battery SOC, charging rate | **Single objective:** Reduce energy costs | Battery specifications | Single |
| [90] | Deep RL | Battery SOC, charging demand, charging cost | **Multi-objective:** Reduce energy costs, Satisfy EV charging demand, reduce emissions | Driving demands, traffic constraints, charging infrastructure | Single |
| [91] | Deep RL, Proximal Policy Optimization PPO | Baseload, Battery SOC, Home SOC, power consumption, electricity price, departure time, renewable energy | **Single objective:** Reduce energy costs | Charging constraints | Multi |
| [92] | Q-Learning | EV position, battery SOC | **Multi-objective:** Reduce grid losses, reduce waiting time, increase EV owner's profit | Battery constraints | Multi |
| [93] | Batched RL, fitted Q-iteration | Time till full-charge, EV availability, flexibility of EV aggregation | **Single objective:** Reduce energy cost | Charging constraints | Single |
| [94] | Q-Learning | Energy Consumption | **Single objective:** Increase customer welfare | Charging constraints | Single |
| [95] | Deep RL, deep Q-network DQN | Battery SOC, available charging places | **Multi-objective:** Reduce charging time and charging fares | Charging availability | Multi |
| [96] | Deep RL, deep Q- | Electricity price, home SOC, | **Multi-objective:** Reduce | Price constraints | Multi |

**TABLE 3.** ***(Continued.)*** **Summary of the RL parameters in the EV coordination problems found in the studied literature.**

| | | | | | |
|---|---|---|---|---|---|
| | network DQN and A2C | renewable energy, power consumption, EV availability, time till departure, battery SOC | energy costs, balance load profile | | |
| | network DQN and A2C | renewable energy, power consumption, EV availability, time till departure, battery SOC | energy costs, balance load profile | | |
| [97] | Deep RL, deep Q-network DQN and DDPG | Electricity price, charging power | **Single objective:** Increase charging station profit | Price, Net power consumption, charging constraints | Single |
| [98] | Q-learning | Electricity price, Battery SOC, time till full-charge | **Single objective:** reduce energy prices | Battery constraints | Single |
| [99] | Adaptive Heuristic-Critic Recursive-Least-Squares AHC-RLS | EV charging demand, electricity price | **Multi-objective:** balance load profile, increase charging station profit | Charging constraints | Multi |
| [100] | ES-MARL Equilibrium-Selection-Based MARL | Battery SOC, number of available EVs | **Multi-objective:** reduce energy prices, reduce customer dissatisfaction, avoid battery degradation | Battery constraints, driving patterns | Multi |
| [101] | Complementary Generation Control CGC | Weather, Number of controllable EVs | **Multi-objective:** reduce energy costs, balance load profile | Power balance constraints | Single |
| [102] | Deep RL, DDPG with Deep Genetic Policy DGP | Weather information, traffic information | **Multi-objective:** Reduce traffic congestion, integrate renewable energy | Power constraints, renewable energy constraints, battery SOC, charging constraints | Single |
| [103] | Batched RL, fitted Q-iteration | EV requested energy, arrival time, time till departure, charging rate | **Multi-objective:** Balance load profile, Satisfy EV charging demand | Parking Places | Single |
| [104] | Q-learning | V2G, TOU | **Single objective:** balance load profile | NA | Single |

energy distribution systems. Some other proposed methods ([58], [63], [72], [82], [85], [94], [96], [99], and [100]) aimed to better control the load profile and used the charging demand as a state parameter, as the load of the system increases when EVs demand electricity.

Methods presented in [55], [56], [59]–[68], [70], [73], [74], [77], [79], [82]–[84], [86], [87], [90]–[94], [95], [98], [100], and [100] used the battery SOC as a state parameter. The objectives of these methods range from cost and profit management to load profile and transformer load management. Battery SOC is a crucial parameter as it indicates at which process is the EV. The EV could be charging, discharging, traveling, or still. Battery SOC is needed almost in all papers as a significant ecological parameter that helps in many decisions. Maximum charging rates of EV charging stations (EVCS), maximum and minimum charging power, ESS, EV consumption at road, and remaining time to charge are all parameters that were used less frequently than the battery SOC. Maximum and minimum appeared several times but were not as frequent as the battery SOC.

Renewable energy sources are monitored in distributed energy systems as some papers were trying to solve the inflexibility challenge of using renewable sources by using reinforcement learning to find a solution to that problem while satisfying the demands of the energy systems. The papers that used it as a state parameter are [61], [69], [72], [76], [86], [87], [88], [91], [96], and [100]–[102]. Emission was used only once by paper [90]. Also, other parameters were rarely used as state parameters, like distance, location, times of arrival and departure, and the direction of EVs.

In conclusion, Battery SOC, EV charging demand, and charging or discharging costs were the most prevalent parameters as state parameters, as they are physically and logically crucial for action selection toward achieving the objectives of the surveyed methods in managing the distributed energy systems. Their recurrence may be the result of the inclination of the authors to control the cost, manage the load profile, and satisfy the charging demand. Other parameters that were less crucial and contributed to less frequent objectives appeared less as state parameters or inputs.

### 2) OBJECTIVES

Although all the methods surveyed were concerned about managing distributed energy systems with EV penetration using an RL approach, the objectives of the solutions presented in the surveyed papers vary. The learning agent takes the necessary actions to accomplish its goal. In other words, the RL agent learns the optimum actions to achieve these goals through assigning rewards for the actions that achieve the goals and penalties for actions that do not achieve the goal.

Most of the studies surveyed in this paper were concerned about the energy cost and the EV charging prices (see Table.3) EV charging prices is a business objective, which makes it a crucial topic to be tackled widely in research. In [57], [62], [77], [79], [81], [86], [87], [93], [54], [56], [98], [102], [89], and [101], the authors presented several single-agent RL-based solutions with the objective of reducing all the energy-related costs in the distributed EV charging systems. The authors of [57], [62], and [79] attempted to develop solutions that reduce energy costs by managing the charging schedules of the EVs and forecasting energy prices to achieve the most cost-effective daily charging schedules for the EVs. Authors in [87] studied the effect of integrating renewable energy generators in the EV charging system with the aim of reducing the energy charging prices.

Also, [101] presented an RL solution to reduce energy-related costs in EV charging systems by trying to reduce regulation costs caused by the imbalance between the generation and consumption sides.

The studies in [54], [56], [98], [102], and [89] provide solutions that target only lowering the end-consumer energy charging price and increasing consumer welfare.

Different studies have different aims, but they all have the same objective of reducing energy-related costs in EV charging systems. On the other hand, some studies aimed to increase the profit of EV charging. Chargers included the EVCS or the energy distribution systems. Authors [61] and [97] presented solutions that maximize the charging stations' revenue in distributed EV charging systems. Likewise, the studies in [58], [60], [63], [64], [65], [66], [67], [68], [70], [71], [80], [84], [88], [91], [94] and [96] presented a likely solution that minimizes the energy costs and increase EV charging profit but with the use of multi-agent RL (MARL) techniques instead of the single-agent techniques. The advantage of using multi-agent or single-agent RL techniques are explained in a later section.

Energy cost seemed to be a reasonable objective as charging services providers care more about profit.

Many studies also aimed to use RL methods to satisfy the EV charging demands while maintaining the energy system needs. In fact, the main goal of using RL is to deal with the uncertainties that emerged from the energy systems' interaction with the numerous EVs. The studies in [73], [74], [78], [82], [83], [85], [90] and [100] presented some RL based solutions with the objective of satisfying the charging needs of the EVs' charging distributed systems. Some studies, such as those in [73], [74], and [83], tried to accomplish meeting the EVs' demands by ensuring that the battery SOC does not drop below a certain threshold for the whole traveling time. The authors of [85] tried to accomplish the same goal by lowering the waiting time of the EVs at the EVCS. Most of these studies relied their solutions on multi-agent RL techniques, while [90] and [100] used single-agent reinforcement-learning-based solutions. Meeting the EVs charging demand is crucial to achieving consumer satisfaction. Another objective that appeared in the literature is achieving a better load profile through balancing the power mismatches and aiming to flatten the power consumption peaks.

The studies [72], [99], and [100] attempted to manage the load on the transformers at the EVCS in order to maintain and ensure that no failures happen in the charging system. The papers [101], [103], and [104] proposed single-agent RL solutions targeting balancing the power between the generation side and the consumption side.

Actually, some papers used more than one objective in their RL models. Other less significant objectives were also considered in some studies like using renewable energy resources in the energy distribution network such as in [72] and maximizing power-saving such as in [99] and [104].

Although all the studies that were surveyed discussed managing distributed energy systems with the introduction of EVs and the challenging uncertainty that appeared with them, the objectives of these studies differed according to what aspect of the problem the authors tried to solve.

### 3) CONSTRAINTS

Ensuring that the learning agent is following realistic and practical environmental scenarios, some constraints must be considered to control the learning process of the agent (see Table 3). Also, constraints are needed to achieve the desired objectives without ignoring other important factors related to the distributed energy system. The nature of the constraints is dependent on the nature of the goals of the algorithms. Since different studies had different objectives for using RL, different constraints appeared in the surveyed papers. As some objectives appeared more frequently than others in these studies, some constraints also appeared more frequently than others.

Most of the proposed methods have used energy network constraints so that the RL agent is guaranteed not to target the ultimately optimum states, but the optimum states that can be realistically achieved by the existing EV charging network. The methods proposed in [54], [57], [58], [60], [63], [67], [70]–[72], [76]–[78], [82], [88], [90], [91] and [94] have all restricted the RL agent with some constraints related to the EV charging network that represents the environment in which it learns. These constraints are needed for the learning agent to avoid unrealistic scenarios. For instance, the authors in [63], [70], and [71] forced some constraints on the transformers' loading to prevent the learning agent from going to network states that include transformers' over-loading conditions. Going through these network states can optimize the desired objectives such as satisfying the demand or reducing charging prices, but these states will harm the charging network and cause many faults in it. Therefore, by using this constraint, the learning agent can be directed towards an effective behavior. In [82] and [94], the utilization of the EV charging stations have been considered in the constraints that are controlling the learning agent behavior. The EVCS utilization was used to measure the loading of the charging network and try to balance the charging demand on the available charging stations to prevent the overloading of particular stations while leaving other charging stations unemployed. This constraint enhanced the charging scheduling and the charging network sustainability. Several methods presented in [58], [60], [71], [72], [76], [77], [88], [90], [91], and [94] used some network limits such as the maximum and minimum rate of charging, the maximum charging station capacity, and the minimum charging station capacity. These constraints represent the physical limitations of the EV charging system, and the authors of these papers claim them to be very necessary to constraint the RL model. In general, network limits are very crucial to be considered as constraints since they represent the physical limits of the system.

Another example is defining some battery parameters like capacity and charging rate as constraints to provide a realistic representation of the battery's behavior in the model. The studies [56], [87], and [100] have used the EV's battery's maximum rate of charging and SOC as constraints in the RL model. These constraints were used to keep the state of charge of the EV batteries always above some threshold. By doing so, the RL model can prevent the depletion of the EV battery. In [60], [62], [67], [79], [84], [98] and [100], the authors also used the state of charge constraint in their models to provide some protection for the EV batteries. Hence, some battery-related parameters have been shown in these studies to have an effective impact on leading the RL agent towards some desired behavior.

Charging demand satisfaction constraints were used in [58], [64], [65], [68], [77], [78], [80], [81], [83], [99], and [100]. These aimed to meet the charging demand of EVs and used the EVs charging demand satisfaction as a state parameter. It is logical to assign a constraint on a parameter you are observing to control the performance of the system.

Other constraints appeared less frequently, like price, power balance, and net power consumption. These constraints did not function as an approach to a realistic representation of the components of the environment, but as an approach to achieve some conditions related to the objective of the authors in different papers. For instance, [65] used EVCS routing topology, driving time, and waiting time, in addition to the charging demand satisfaction constraint. Price was used as a constraint in [96], [97], and [101]. Price is a target to be achieved through the algorithms, using it as a constraint might not be of significant need in most cases, as the agent is already learning to find the optimum prices. The price constraint might have been a second degree of controlling the energy prices. Power balance constraints were used in [101] and [102] only, as balancing the power usage was not a frequent objective in the surveyed paper. Renewable energy constraints appeared in [69], [72], [76], and [102]. However, only [76] aimed to manage renewable energy sources, while the others were concerned mainly about the cost.

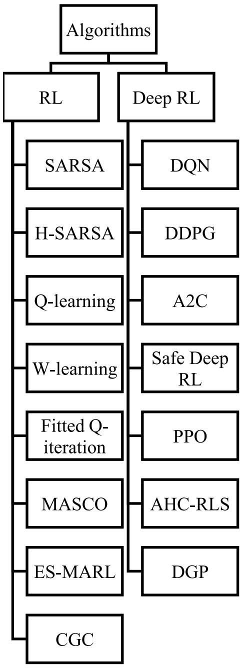


**FIGURE 7.** Model-free RL algorithms used for the EV charging coordination problems.

Constraints by nature are a part of the solution, and their relationship with the objectives and behavior of the system is important because it can lead to the failure of the model if not well selected. Therefore, constraints may be used as an approach to control the behavior of the learning agent to achieve a realistic solution to the defined problem or to achieve a desired objective for the algorithm.

### 4) ALGORITHMS

Several algorithms were used to solve the problem of EV charging systems with the uncertainties (see Fig.7).

The methods used in model-free RL are classified RL, or Deep RL which combines RL with deep machine learning. Most papers used Q-learning, or the deep extension of it, DQN as a method to solve the RL problem. Other papers tried to solve some issues that happened to exist with Q-learning with other techniques, like W-learning. Some papers did not consider Q-learning and used SARSA, State-Action-Reward-State-Action, to solve the learning problem. In this section, we provide an explanation of each algorithm used in the surveyed paper and the reason for using it.

To develop an RL model that can achieve a certain objective for the EV charging system, the authors of [62] considered modeling the EV charging problem as a Markov Decision Process (MDP). The MDP is a mathematical model that is usually used in problems involving achieving optimum decisions.

In [62], the authors suffered from this drawback as they could use only three inputs and discretized them to very limited levels, which affected the representation of the charging system environment. Another drawback is also the discretization of the possible actions, which force the agent to select an action from the pre-defined actions set; consequently, it limits the model from achieving the optimum actions.

In [58], the authors considered an enhanced version of the SARSA algorithm which is the Hyperopia SARSA (HAS). The HAS learning algorithm is the same as SARSA but with the exception of representing the Q values of the state-action pairs as a linearly approximated function instead of a table of values. This enhancement enabled the study to use continuous spaces for the states and actions, which removed the disadvantage of the restricted environment representation as a result of discretizing the state space. This enhancement can also help the agent to take better actions through the use of continuous action space. However, the algorithm still suffered from the remaining conventional SARSA drawbacks.

Most of the studies in the literature, though, used the Q-leaning as the learning algorithm for their RL-based solutions. All of the studies [54], [60], [61], [66], [68], [74], [73], [77], [80], [81], [83], [86], [87], [92], [94], [98], [100], [101] and [104] have used the Q-leaning algorithm. The Q-leaning algorithm is an advanced version of the SARSA algorithm. It defines a set of state values (V), a set of actions (A), and a set of state-action pairs (Q). Moreover, it follows the epsilon greedy algorithm for picking the actions and accordingly updates the state values and the state-action pairs' values. However, the Q-learning algorithm considers the whole future trace and guarantees reaching the absolute optimum policy.

Moreover, the Q-learning algorithm still has issues with high-dimensionality and the discretization of the state-space and action-space. The curse of dimensionality is the inability of using many input features in the RL model. The problem of the discretization of the state and action spaces can be solved by the integrating function approximation with the Q-learning. The resulting algorithm from integrating function approximation with the Q-learning algorithm is called the W-learning algorithm, which has been used in some papers like [70], [71], [72], [75], [78], and [82]. However, this can affect the algorithm badly to the extent that it can make the algorithm diverges and not converges to any policy at all. That is why most of the studies always prefer not to integrate function approximation with the Q-learning algorithm even at the cost of discretizing the input features. This makes the Q-learning algorithm suffer from both discretization and dimensionality problems.

A deep version of RL was introduced in some papers to deal with the high dimensionality of the state space and the action space. The DRL approach represents an extended version of Q-learning, where the learning agent uses deep neural networks (DNN) as the function approximator. The type of DNN used is deep Q-Network (DQN), which returns a set of action values for each state. The papers that used deep RL and DQN are [53], [61], [63], [67], [82], [86], [88], [89], [93], and [94]. Two important features of the deep RL and the DQN algorithm are the target network and experience replay. The target network is of a similar structure as the used DNN, but the parameters in the target function are updated in a slower manner, and the values do not change except at the update time. Experience replay replaces the immediate state transitions by storing them in a buffer and sampling them uniformly to update the network. Furthermore, the introduction of the deterministic policy gradient DPG to the DQN method extends it to continuous control. Therefore, the deep deterministic policy gradient DDPG is used along with the DQN approach.

In [67], the DQN and DDPG approaches were modified, and a control deep deterministic policy gradient (CDDPG) was presented by the authors. They used a type of RNN called LSTM, long short-term memory, to extract state features from the environment and DQN to select and execute the actions. They stated that the behavior of their algorithms included three stages interaction, training, and testing. This approach relies on experience replay and target network features of deep reinforcement learning. Another paper that improves the DQN and DDPG approach is [97]. The authors presented prioritized experience replay (PER), which gave higher priority to experiences that dealt with higher absolute TD error, deriving better policies and learning faster.

A batch RL method that was used in several papers is fitted Q-iteration, where the state transition dynamics are learned from a batch of sampled transitions. On-policy methods like SARSA, update the policy used in the action selection agent. However, fitted Q-iteration, like Q-learning, is an off-policy value iteration method, learns the optimum policy through iterations till the best action is selected. The agent's behavior is independent of the updated policy. Hence, it does not follow a greedy policy to select the actions. The optimum policy for action selection is based on experiences gained by deploying a random policy then observing the transition dynamics. The papers that use fitted Q-iteration and batch RL are [56], [57], [76], [79], [85], [93], and [103].

All the RL models that have been discussed so far are explained in the context of an RL class called the single-agent reinforcement learning. However, some studies have used these RL algorithms in another context which is another class of RL called multi-agent reinforcement learning (MARL), which has been used as well in the literature of RL-based solutions for EV charging systems. The advantage of the MARL over the single-agent RL is that it can allow for much more complex environments that have high feature dimensions, high action dimensions, and continuous space for both states and actions. That is because the MARL algorithms involve multiple agents interacting together in the environment. So, MARL algorithms can overcome the problems that have been encountered in the single-agent RL models which are the dimensionality problem and the discretization problem. However, MARL is not used so widely in the literature since its over-complexity and the non-stationarity of the algorithms. The over-complexity of the MARL is a consequence of the use of a usually high dimensional complex environment and the need to train several learning agents simultaneously, which leads to the complexity of the theory of the algorithm and the expensive computational cost of its implementation. The non-stationarity of the environment is also another problem that arises in the MARL because of the use of several agents learning independently in the same environment. So, every learning agent has its own perceived version of the same environment. The problem arises because this agent's version of the environment can never be fixed, as many other agents regularly alter the state of the environment. In other words, the state of the environment changes based on the joint actions taken by all the agents, not the actions of a single agent. Hence, the non-stationarity of the environment is a crucial challenge that has to be addressed when using MARL models. The authors of [99] and [100] have used MARL algorithms in their RL models. In [99], the authors used ES-MARL algorithm, while in [100], the authors used AHC-RLS algorithm. Actually, both of these MARL algorithms are based on the concept of deploying multiple agents in the environment. Each agent has its own reward function, and it learns independently from the other agents. However, there are some non-avoidable interactions among the learning agents. In these studies, the agents' interactions have been dealt with through sharing the complete information of the learning agents with some tweaks to lower the computationally cost of this information sharing. By doing so, the algorithms have been guaranteed to converge, and with the right representation of the environment, the algorithms have been guaranteed to converge to an optimum policy for the problem.

As presented above, Q-learning, and in some cases SARSA, was used to solve the RL problem of managing distributed energy systems. However, concerned by the high dimensionality of the action and state spaces, other techniques were proposed. The most used one was DQN along with DDPG, followed by W-learning. DQN integrated with DDPG represented a good solution to the high dimensionality problem while extending the learning model to continuous models. Of course, the choice of algorithm is always related to the type of problem the authors of the papers tried to solve, see Table 4.

**TABLE 4.** Pros and cons of main RL algorithms.

| Algorithm | Pros | Cons |
|---|---|---|
| SARSA | - Simple<br>- Guaranteed to converge to an optimum policy | - Does not consider future actions<br>- May converge to a local optimum<br>- Limited number of states |
| Q-learning | - Considers future actions<br>- Converges to a global optimum | - High-dimensionality issues<br>- Discretization of state and action spaces |
| W-learning | - Solves the discretization issues | - May not converge to any optimum policy |
| Fitted Q-iteration | - Converges faster<br>- Continuous state and action spaces | - High-dimensionality issues |
| DQN | - Solves the high-dimensionality issue | - Discretization of state and action spaces |
| DDPG | - Continuous state and action spaces | - Assumes experiences are of the same priority |

## D. THE APPLICATION OF RL IN EV CHARGING MANAGEMENT

A summary of the RL approaches used in managing energy systems containing EVs is presented in this section. The application of reinforcement-learning solutions in smart grids [26] is inspired by their results in the complicated scheduling problems involving EVs [25]. The recent reviews in model-based EV charging coordination show that such algorithms are hindered by (i) the need for a precise system model, and (ii) cannot be a generalized solution. We categorize the RL applications based on the location of the controller such as is in residential charging, centralized or distributed, and if it is real-time or day-ahead planning (see Fig.8).

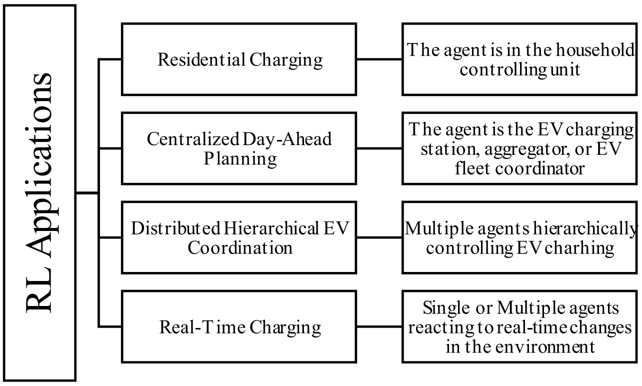

RL Applications
Residential Charging
The agent is in the household controlling unit
Centralized Day-Ahead Planning
The agent is the EV charging station, aggregator, or EV fleet coordinator
Distributed Hierarchical EV Coordination
Multiple agents hierarchically controlling EV charhing
Real-Time Charging
Single or Multiple agents reacting to real-time changes in the environment

**FIGURE 8.** Types of RL application systems for EV charging management.

### 1) RESIDENTIAL CHARGING

A study in [54] proposed using RL to learn the individual energy consumption in a household then use the information for EV battery charging management to maximize the total individual welfare which is obtained from household consumption and EV charging.

Individual EV user behavior profile is created to model the customers based on statistical data using a bottom-up design approach [105]. The EV user's driving distance is obtained from mobility data and the distance per day required is related to the social groups part-time employee, full-time employee, students, and retired unemployed people, and the driving behavior is related to weekends/weekdays activities. The model of EV customer gives the charging demand according to the inputs given used to simulate the driver's behavior. A stochasticity factor is considered based on probability from historical data which considers the uncertainty of social group, activities thus daily driving distances. EV charging is scheduled optimally with respect to individual maximum welfare. RL is applied to learn household consumption and the Q-learning mechanism algorithm reaches the learned policy corresponding to the individual household consumption over a daily horizon. The learning mechanism learns the household consumption trend through iteration over the states which represent consumption levels. The states are the consumption which are levels of 1 watt, i.e., discrete values. The agent chooses the policy that results in the highest utility from each state corresponding to the highest reward. Rewards are related to the utility function where the higher the utility, the closer the state is to the actual household consumption. Then, after learning the consumption of each individual, the agent adjusts the EV charging to maximize the total individual welfare, which is obtained from household consumption and EV charging.

The study in [71] presents a residential EV charging coordination based on three different policies to avoid overloading the transformer, ensure minimum desired charging achieved, and charge during off-peak loading. Five scenarios are considered, baseline with no control where EVs start charging once they arrive home. In the second scenario, agents have information about the load to allow charging during off-peak with high reward and is panelized otherwise. The third scenario investigates periods for charging during medium and high peaks when the off-peak period is insufficient to charge the EV demand. In order to enable the agent to learn this scenario, information about the predicted load for the next 24 hours is provided in the state space. The fourth scenario considers that all EVs need a complete charge for a considerable mileage. In this case, the agents learn to spread out their charging periods while considering the transformer limitation. The fifth scenario enables the agent to charge based on the current load and predicted loads.

### 2) CENTRALIZED DAY-AHEAD PLANNING

The authors in [56] solve the EV scheduling problem based on day-ahead and following predicted electricity prices. Their goal was to reduce the individual PEV charging cost in the long-term. The RL algorithm utilizes a set consisting of historical samples of transition to learn the optimal policy leading to lowering the charging cost. The authors modeled the problem of calculating the energy used in charging as a decision-making problem on a day-to-day basis, with unknown transition probabilities. To predict energy prices, the authors used a Bayesian neural network, while using historical data to train the RL program. Consumption patterns and plug-in and plug-out states in optimization linear programming are used to select the best action in every scenario. The MDP is modeled using continuous state-space and daily timesteps. Therefore, the aggregator can calculate the energy needed for buying from the energy market for the next. The aggregator controls the charging power of the EVs to minimize imbalance costs between day-ahead purchased power and the actual load. Therefore, the aggregator has to custom itself to its corresponding fleet of EVs flexibility to minimize the cost difference.

An electric vehicle fleet dispatch solution for the problem of ride-hailing services is presented in [66] using RL with the aim to bring the waiting time of the customers, EVs' operational costs, and electricity cost to a minimum. They use decentralized learning combined with centralized decision-making for finding a solution to the MDP-modeled EVs fleet problem. In the training process, individual EV experiences from interacting with the changing environment are used to train the common state-value function approximator and its parameters. Then, in the centralized decision-making process, the coordination of the EV fleet is solved as a linear assignment problem by using the estimated parameters state-value function in the learning stage. Therefore, the scalability of the algorithm is improved, and the results show better performance compared with benchmark algorithms concerning the operating costs of the EVs, as well as the customers' time of waiting.

Another day-ahead scheduling plan to charge a fleet of EVs is presented in [57], and it considers addressing flexibility by a heuristic charging control scheme. The study aims to produce a day-ahead plan for the charging consumption of the EV fleet. Batch RL is used to predict a day-ahead plan for a cost-effective charging consumption. The resulting behavior from various factors is used by the learning agent to produce the plan. Furthermore, controlling the actions of the whole fleet of EVs instead of every single EV reduces the size of the state and action spaces. The objective of the charging strategy is to allow for the aggregator to manage an EV fleet demand schedule by purchasing the energy in the day-ahead market. In day hours, the aggregator receives charging requests from the EVs and decides on the power for each EV (control power) based on the objective to balance the difference between the requested demand and purchased energy. A dispatch algorithm obtains the state information and schedules purchased power, then calculates the control power values. The states considered are minimum charging power, maximum charging power, requested energy, and departure time. The benchmark used to test the algorithm was a multistage stochastic-programing solution. Then a heuristic division strategy based on historical data is used to retrieve the individual control actions back from the collective control ones.

To overcome the problem of continuous spaces, and to provide a generalization of different observed values, fitted Q iteration in [106] is used instead of temporal difference learning [107]. The solution is evaluated for a small company with an EV fleet of 15 EVs where there are 4 EVs in the morning shift, which starts from 6:00 till 14:00. 8 EVs work in the day shift, which starts from 9:00 till 17:00. Therefore, 3 EVs work in the afternoon shift, which starts from 12:00 till 20:00 [108]. The scenarios were formed by the assumption of time of departure, time of arrival, and wanted amount of charging energy. The amount of energy requested by the EV is obtained according to an artificial probability distribution to provide the variability needed for the benchmarking of the presented solution. The study did not consider comparing different pricing mechanisms [109] used for the charging flexibility of EVs by motivating the consumers using incentives.

### 3) DISTRIBUTED HIERARCHICAL EV COORDINATION

A distributed, multi-agent EV charging coordination system considers the network impact is presented in [110]. The agents considered are based on the Microgrid concept [111]. Electric Vehicle Supplier/Aggregator (EVS/A), Regional Aggregation and Microgrid Aggregation Units, (RAU) and (MGAU), and Vehicle Controller (VC). The control method is based on Nash Certainty Equivalence Principle to consider network impacts. The controlling strategy of EVS/A aims to maximize the charging profit by valley-filling. The RAU agent optimizes network operation, while the VC agent tries to satisfy the charging demands. MGAU agents treat the network and grid violations using RAU pricing policy [111]. Game-theoretic methods are used to solve the pricing policies [112] The proposed charging management system is tested for its efficiency using an urban network realistic simulation for different cases. The study considers dumb charging, dual tariff, uncouples EVs, and weekly coupled EVS, and for both cases summer and winter. The results are of the same range of quality as the centralized approach, but requiring less communication and less computation power. The better hierarchy improves the scalability of this approach. The proposed MAS system is written using a Java developing agent (JADE) [113].

### 4) REAL-TIME CHARGING

The random nature of the EVs in the times of arrival and departure, rendered the traditional day-ahead learning algorithms applicable, due to their indifference to the real-time behavior of EVs.

In [55], the authors presented a solution to calculate the optimal policy of charging schedules using deep reinforcement-learning, or DRL. The objective of the presented solution is to minimize the EVs' overhead, such as charging fees and time. The deep learning approach uses the TOU tariff for the algorithm while using the charging availability and parking places as constraints. The constant speed movement of each EV is independent of the others on two-way roads, while the EVs ask for fast charging services. The online collected information at a certain moment are battery-state-of-charge, location, speed, and direction of each individual EV. The highest priority for charging is based on first-come-first-served (FCFS). In the FCSs, the rate of charging is assumed to be fixed for each plug-in charger. The TOU mechanism is applied where the electricity price varies on the same day, having its own flat, peak, and valleys. The charging price is calculated momentarily when the EV charging starts, it is assumed that the users would not stop the charging for another time with a lower price of electricity. Deep RL is an extension of Q-learning to solve the dilemma of the high dimensions of the action and state spaces. The learning agent uses DQN, which is a DNN, in the role of function approximator. Considering more environmental factors like the EVs changes in the real-time velocity is recommended by the study.

An online strategy to obtain optimal energy prices and charging schedules, which takes into consideration the random behavior of EVs in arrival and departure, proposed in [58] is based on the RL algorithm solved with Hyperopia SARSA-based approach (HAS). It is a model-free based decision model that is based on stochastic models of assumed uncertainties in future events. The control mechanism tackles the problem of real-time variation in the state-space and action-space that results from the randomness of the EVs' arrival and departure times. The state-action function is presented by linear approximation. The algorithm learns only from the observations of past events, such as the arrival time and the departure time of the EVs, and the prices of electricity to charge the batteries. The time in the system model is divided into slots. During every time-period, the charging station determines the charging rate based on previous observation including departure time and charging demand, and the decision then is affected by the remaining charging demand at the station left for future EVs. Therefore, the problem is an MDP problem that aims to satisfy the charging demands of EVs before their departure by optimizing the overall rates of charging. The proposed algorithm is experimented with using different EV load profiles, namely emergent, normal, and residential from a historical database. Results are compared with Robust Simulation-based Policy Improvement (RSPI) [114]. The presented Hyperopia SARSA-based algorithm (HAS) is also compared with sample average approximation (SSA) and greedy policy.

Authors of [60] presented a real-time algorithm based on V2G control services for EVs to decide whether the EV should charge, discharge, or regulate the frequency. It is a novel V2G control algorithm that considers the uncertainty of electricity prices as in the demand response algorithm [115]. The algorithm can learn from past experiences and automatically adapt to the unknown pricing and makes hourly optimal control decisions for V2G control. MDP is formulated after one EV's point of view. They maximized the owner of EV's profit in the time of parking. To achieve this, they used Q-learning to control the operation after modeling the electricity price with a two-state Markov chain, while not knowing the probabilities of transitions. The electricity price is calculated in an hourly manner. The EV arrival time is assumed to be 6:00 p.m. every day with a SOC of 40%, while the departure time is assumed to be 8:00 a.m. with a SOC of 70%. With the absence of coupling constraints, the state-space and action-space are if small size, as MDPs are formulated for the individual EV. The control actions are coordinated for creating a collective day-ahead schedule.

Authors of [61] proposed an online RL-based charging application station automated electricity distribution system (SAEDS) to increase the profits of a single EV charging station connected to a renewable energy source (RES). The learning agent uses Q-learning to determine the EVs' dynamic charging amount of energy, based on the prices of electricity and hourly generated renewable energy. The algorithm presents a reward as the station's profit calculated by the sum the station earns at the current hour of the day. The station has k slots, a max number of vehicles charging or waiting, time to leave (TTL), state-of-charge (SOC) the energy from the RES, and the extra price of energy bought from the grid. The non-homogenous Poisson model was used for the arrival time of the vehicles, while historical data was used to estimate the number of EVs arriving at certain times. A normal distribution with different parameters for the different day hours is used to obtain TTLs of arriving vehicles. The authors introduced a sigmoid function to present user type based on the willingness to pay depending on the difference in price between their initial SOC and the SOC they want to reach. The algorithm outperforms the solution of using uniform distribution over each possible decision, on every step, the increase in the station's revenue is in the range of 40–80%.

The previous studies in the literature proposed energy scheduling on short time periods such as hourly-based scheduling, whereas battery charging is scheduled on a daily basis. Additionally, energy prices are given by the utility over the entire scheduling horizon ahead of schedule in some cases, and in other cases, they are unknown [60]. In [62], the authors considered the price uncertainty by combining the day-ahead energy prices, which is known, with the second day-ahead energy price, which is predicted, to develop EVs charging schedules. The objective of their algorithm is to minimize charging costs while ensuring the full charging of the EVs. The authors used state–action–reward–state–action (SARSA) to learn the changing patterns of the price of electricity, solving the problem of charging. Also, they assumed that driver's daily patterns are known. They also modeled the charging/discharging schedules as a constrained Markov decision process (CMDP) and used safe deep reinforcement learning (SDRL) to solve the problem. One year of historical data was used for the prices of electricity, while a truncated normal distribution was used to estimate the battery energy, the times of arrival and departure, and the behavior of driving.

The aforementioned studies provide EV schedules using RL approaches based on the driver's patterns like arrival and departure times, and also the SOC of the battery at the arrival of EVs. The values of these parameters were fixed or estimated using normal distribution or real driving patterns or even modeled using Poisson stochastic processes. Unfortunately, these models fail to provide sufficient information to predict the behavior of the several users at a single charger, rendering it impractical for real-life applications. The authors of [63] generate the arrival time and charging duration using kernel density estimation (KDE) and then solve the scheduling problem using model-free RL algorithm DQN. Furthermore, the proposed EV management deep RL algorithm aimed to reduce the charging cost of EVs, while at the same time increase the reliability of the grid. The approach considers charging and discharging at one public charger which can be used by several users instead of only one charger per person. The action space presents the charger's charging/discharging levels as a discrete space (−4 kW, −2 k kW, 0 kW, 2 kW, 4 kW). The authors assumed users required full-charged batteries. The remnant energy in the battery at the start of the charging was estimated based on the end-time charging and the efficiency of the charger. The results of the proposed method, random variable-RL method, and fixed-RL method are compared with the cost of the unscheduled charging (for specific site A) which are 69.7%, 74.6%, and 75.7% respectively.

There is no guarantee that an EV has a fully-charged battery at its time of departure set in a specific charging schedule because of the random behavior of the EVs in arrival and departure times. The authors in [64] address this issue by considering the EV problem as a constrained Markov Decision Process (CMDP) and solves the problem with constrained policy optimization (CPO) [116]. They propose a real-time strategy for charging and discharging the EVs that tackles the random behavior of EVs in arrival and departure, prices of electricity, and battery remnant energy. A constrained optimal charging/discharging schedule is obtained using SRL which is a model-free approach and uses a deep neural network (DNN) without the need of setting of penalty coefficients manually as in DRL. The states at each time step are the battery energy and the past 24 hrs. electricity price. Furthermore, the action is set to be the charging and discharging quantity and it is a continuous variable. The reward during charging action is presented by the electricity cost but as a negative quantity. On the other hand, the revenue from selling electricity is the reward while discharging. The buying cost and the selling cost of electricity are assumed to be the same. The optimal policy is assumed to have the maximum discounted return. The state parameters of the system are the network inputs. The extraction of features from the information in the system states is the role of the policy network. The output of the policy network is a normal distribution, including its mean values and logarithmic standard deviations. Then, the action of the EV charging or discharging is generated by the policy after sampling the normal distribution. State values are produced by a value network after the extraction of state information. The value and policy networks are of similar architecture. The proposed method is compared with baseline approaches Deep Q-Network (DQN) and Deep Deterministic Policy Gradient (DDPG) with a penalty term in the objective. Also, the approach is compared with the deterministic optimization problem solved using SCIP [117]. The algorithm code is written in Python3.6, especially using TensorFlow1.12, as it has a deep learning package.

A control system proposed in [59] addresses the voltage security by maximizing the operator of the distribution system's profit while considering the power system's physical constraints. Their approach analyzes how uncertainties affect the charging strategy by deploying the DDPG-based RL technique.

In [64], the EV user's perspective is considered and the effect on the user's decisions on electricity prices is not considered. It is worth investigating the effect of adjusting the electricity prices when new peaks are developed on the EV's learning.

Different from the previous studies, the authors in[65] proposes a charging navigation application aimed to minimize the EVCS charging cost and the EV traveling time without requiring the uncertainty data of a prior system. An MDP is used to model the problem with the transition probabilities unknown, while the state features were extracted by a deterministic shortest charging route model (DSCRM) from stochastic data. The system states are constructed according to the road velocity, the charging price of the EVCS, the time of waiting of the EV, the location of the EV, and battery SOC at every timestep. The reward is realized by the waiting time and charging cost at the EVCS, the charging cost is assumed constant for all charging groups in this study. The action by the EV driver determined the EVCS and the corresponding rout.

The charging strategy in the previous studies did not consider the power system limits. While the study by [70] propose simultaneously minimizing the energy cost and transformer loading. It proposes a MultiAgent Selfish COllaborative architecture (MASCO) which is a multiagent multi-objective RL architecture used to develop a heuristic controls strategy to schedules the EV charging at appropriate times. The RL framework aims to optimize for high battery level, minimum price paid, and minimize transformer overload. The proposed policy is compared with the always charging policy (ACP), ransom policy (RP), and DWL-DWL, and the experiment was implemented in BURLAP.

### *E. NOVEL RL ALGORITHMS*

A study by [67] uses RL to learn the optimal strategy to satisfy the EV's battery energy requirement and to minimize the charging expenses at the same time. The method proposed is to control deep deterministic policy gradient (CDDPG), which can solve continuous problems in real-time [26]. This approach is superior to the deep QL-based approach, which is limited to discrete estimates only, and also the DDPG approach.

The uncertain nature of the parameters, such as energy prices, can be addressed as an MDP [118]. The (MDP)-based algorithm assumes known and stochastic future events distributions.

A novel multi-agent charging strategy in [68] considers EV users participating in the electricity market by deploying V2G technology for selling or buying electricity to the grid without any aggregator or central agent. The method uses a Q-learning algorithm where agents buy energy and sell it aiming to minimize costs while always satisfying the trip energy needs. The algorithm also considers the behavior of owners of the EVs and the uncertainties involved. The authors assume each EV as an agent and model the variable times of arrival and departure and energy required during the trip as probability density functions (pdf) of Gaussian distribution. The study considers three scenarios, fixed-rate strategy, TOU and dynamic tariff strategy. Stochastic driving behavior was simulated considering only one trip per day for each vehicle. At first, the agent determines the energy required for the trip and the market bidding price of electricity. A penalty is assigned when the algorithm determines an insufficient amount of energy for the trip at departure. Then, EV owners should determine the benefit of selling electricity to the market. Subsequently, the RL algorithm determines the policy based on the highest reward calculated by subtracting the revenue from energy sold from the buying cost of energy. The effecting factors of time-independent environments are electricity prices, driving patterns, and random energy consumption for each trip. The system considers 500 agents to simulate the three scenarios, fixed-rate strategy, TOU, and dynamic tariff, and the average cost results are 5.5 €/day, 5.3 €/day, and 4.8 €/day, respectively.

An energy management system for EVCSs with solar PV and an energy storage system (ESS) is proposed in [69]. The novelty of this study is that the scheduling solution is computed in a distributed manner while handling time-varying data related to the charging of EVs, such as the amount of energy charged in the ESS, the charging price, and the total demand of the EVCSs. The proposed scheme is based on a multi-agent RL framework with solar PV and an energy storage system (ESS) at each EVCS. For policy choosing, the rewards are related to the EVCS operation cost of charging, discharging, and loss caused by overcharging during the charging process.

Authors of [72] applied a multi-agent RL control for maximizing renewable energy use with EV charging management by considering that each load is a separate device in charge of its own schedule. The state information is the generated power from a renewable source (wind power), load power, and the charge energy. Three policies are proposed in this study, the renewable energy policy which can enable EVs to learn charging only during the periods where there is sufficient load that enable EVs charging, and it is the difference between transformer load and the maximum specified load. The second policy is the battery charge policy which gives a positive reward to agents every time a battery charge is charged to 60% and every time the charge is increased. The baseload policy provides agents with the predicted load of 24-hrs which is related to the energy price directly. The results show that the algorithm enables the EV load to follow the wind generation patterns up to 35% of the overall energy.

State-of-the-art algorithms usually concentrate on single EVs or central aggregate controllers to control an EV fleet through multiple steps which are aggregate load decisions and sending translated decisions to individual EVs. The authors in [85] propose a coordination approach that does not need a heuristic algorithm but instead provides coordination for a set of charging stations that decides which cars to charge to meet the schedule of a target load (i.e., load flattening) in the RL framework. They formulate a scalable MDP that can work on a group of EVCSs that do not share the same characteristics. The optimal chagrining policy is learned by batch RL and fitted Q-iteration. The problem is formulated by setting a maximum number of charging stations for different EV groups and charging levels, the arrival and departure of EVs are considered dynamically over discrete timeslots. The variables considered are the arrival time, the time remaining for departure, energy requested and the charging rate of the EV.

RL approaches determine the charging schedule quality by using action-value functions. Estimating the best action-value function differs from one approach to the other. A Q-table of discrete estimates the action-values of the charging schedule in [80]. However, discretization cannot represent a large number of action-sets required for real-world application. In [58], the authors use a linear approximator to estimate the action-value function. However, it shows a limitation in handling real-world scenarios which are of nonlinear action-value function. Non-linear approximators are used to estimate the action-value function, as in [56]. They use kernel-averaging regression, which requires manual selection of the kernel function with carefully designed parameters.

In [84], the authors used deep reinforcement learning, as they proposed a function approximator for the actin-values to control the EV real-time charging, LSTM network predicts electricity prices. The proposed approach has an architecture of two networks. A Q-network estimates the action-value function, while a representation network represents the prices of electricity. The information presented by the states are EV availability at home or not, the remnant battery energy, and the price of electricity for the past 24 hrs. The actions presented in the problem are charging or discharging power and the problem is constrained by the maximum and minimum power allowed for charging and discharging for the discrete action space. As for rewarding an action, a reward is defined as a positive value of the charging cost and a negative value of the same cost if discharging action. The authors present the term "range anxiety" to the uncharged energy which introduces insufficient energy to reach a required destination by the user. As a result, a penalty proportional to the amount of uncharged battery energy is presented in the RL problem. To develop a schedule, a deep neural network is used. The random variables of EV arrival, departure, and state of charge are modeled as truncated normal distributions. YALMIP solves the problem of the charging and discharging schedules. The amount of energy saving depends on the driver's preference presented by the anxiety coefficient in this study. For instance, a large anxiety coefficient can be interpreted as a small anxiety range, which results in a large charging cost.

## V. LESSONS LEARNT

### A. RL LIMITATIONS

In all the papers surveyed, RL methods had proven to be outperforming the traditional distributed energy management methods and continuously improving themselves. Whether the RL methods were tested against a benchmark, historical data, or sophisticated simulation models, are proven to be superior to their predecessor methods. In the algorithms section, we surveyed how several papers used different RL methods to solve the distributed energy management problem.

Starting from the primitive SARSA, passing by Q-learning, reaching DQN with DDPG, all these methods proved to be superior to the previously surveyed literature in their corresponding papers. However, an interesting pattern appears after further analyzing the nature of the algorithms used. SARSA was used in [58] and [62]. Q-learning seemed to provide better policy selection and was adopted by the majority of the papers. However, the curse of dimensionality problem was introduced and needed a solution.

Thus, W-learning and deep RL were proposed by several papers. W-learning was observed to have a high probability to diverge, even though, it can extend to continuous models. Therefore, the majority of papers tended toward deep reinforcement learning. DQN was the method used to solve the high dimensionality problem using DNNs. It proved to be performing better than Q-learning in the surveyed papers. DQN was integrated with DDPG to enable the use of continuous models. Therefore, DQN integrated with DDPG solves most of the problems that arose with the introduction of RL to distributed energy system management. Thus, we can conclude that deep reinforcement methods like DQN can lead to better results.

To test the performance of proposed RL methods, the majority of the surveyed papers used historical data and statistics, while other papers relied on sophisticated simulations. Of course, both of the testing schemes are not as promising as using the models in real life. However, many papers assigned constraints to ensure the practicality of their methods. Others also induced parameters to imitate real-life systems. These testing methods avoid applying the RL methods in real life to obtain results and serve as a safe area for experimentation. Using new methods to solve real-life problems might introduce risks of underperformance, failures, and custom dissatisfaction. Therefore, predicting the behavior of the RL methods on historical data and sophisticated simulation test cases might give the researchers an idea of how the algorithms might work in real distributed energy systems, and thus, the decision of using the novel algorithms in real systems, if available, may be based on more solid information.

Nevertheless, that brings us to the point where we discuss the application of RL methods in real distributed energy systems. Although RL methods are proving to be superior to the traditional ways of managing the distributed energy systems with EVs, this is all a theoretical approach to solve the problem. The RL approach, and the machine learning approaches, in general, are less computationally heavy than using optimization and centralized control of distributed energy systems and they converge faster to an accepted suboptimal solution. Still, they need serious computational power, which would cost the distributed energy systems to place machines in every EVCS. Even if a multi-agent approach is used, a hardware design problem would present itself, as choosing the hardware to be battery-efficient with the desired performance cannot happen arbitrarily. In conclusion, RL methods are promising but still need improvement to be integrated into real-life distributed energy systems.

### B. RL AND THE FUTURE SMART GRID

The smart grid's complexity is gaining attention worldwide, and because it requires communication and extensive data collection, RL techniques can enable smart grid development and future technologies. RL is a smart tool that enables applications such as energy management, EV charging coordination, vehicle to grid (V2G), Vehicle-to-vehicle (V2V), Vehicle-to-Building (V2B), and Vehicle-to-House (V2H). These technologies are the most recent adopted applications deploying the smart grid [111] and [6].

There are utility advantages of V2G where injecting electricity back to the grid can reduce peak loads and electricity generation cost. V2H can reduce the house's energy bills or enable the vehicle to become an emergency power source from an individual perspective. The same concept is applied to V2B, where a firm or company can benefit from this technology. Other than buildings, V2V enables vehicles to charge each other. Consequently, it will reduce range anxiety, which is an effecting factor on EV adoption [6].

Adopting smart charging is challenging with the uncertainty and complexity of power networks [1]. The drawback of smart EV charging or discharging is the battery degradation, uncertainty of EV arrival or departure, user behavior dependency, and lack of data about the environment or system. RL utilization enables smart charging, load prediction, demand response, renewable generation prediction, Microgrid, and dynamic pricing of electricity, which all fall under smart grid deployment [1]. In summary, model-free RL with smart grids enables EVs, renewable energy sources, and storage systems to share an infrastructure that enhances both economy, power grid, and user satisfaction but present-related research is still at their initial stage. Indeed, RL will have great utilization potential for smart grids in the future, but at present, the relevant studies are still at their initial stages.

### C. RECOMMENDATIONS FOR FUTURE RL APPLICATIONS

In this section, future prospects for EV energy management based on RL are discussed. The current EV management approaches share the need to adapt to some issues. First, they need to tackle the problem of heterogeneous agents and how agents react with each other. Also, the models developed for the distribution infrastructure and their interaction with EVs are either impractical or suggest changing the existing infrastructure. Furthermore, the currently proposed approaches mostly consider one type of energy tariff and cannot be applied for other tariffs. Moreover, there is an unrealistic assumption among the existing approaches that the users know the precise amount of energy they consume each day.

The future steps in deploying reinforcement learning for EV charging management starts with applying the successful RL algorithms in real energy management systems. Then, the integration of the smart grid with the EV management systems is the next step. Therefore, the optimized control of energy management systems will cover a wide range of applications and grow more complicated. Furthermore, multi-agent deep RL approaches will be developed to allow the collaborative interaction between the charging stations and the numerous EVs.

As it has been shown in the last few sections, RL has a great potential in contributing considerably to enhancing the distributed EV charging system applications. The literature in this field is greatly rich and promising. However, the integration between RL and EV charging system applications is still emerging and facing some deployment challenges as discussed in the previous section.

Some researchers, as in [75] and [82], have used single-agent reinforcement-learning approaches in their EV charging related applications through the deployment of the developed RL applications in the EVCS, but most of them have suffered from the limited representation of the environment with complex relationships that the single learning agent can process poorly. Other researchers as in [99] and [100] have used multi-agent reinforcement-learning approaches in their EV charging related applications through the deployment of the developed RL applications in the numerous mobile EVs and having them cooperatively take the optimum actions, but most of them have suffered from the EVs complexity and the poor computational structure of the EV charging system in general.

In recent years, a new technology has been proven to be so promising in many of the real-life distributed system applications, which is fog computing. Fog computing is a computing architecture that positively enhances the computational power of any distributed system. The idea behind fog computing lies in the concept of vertical sub-layering the system. Due to its proven efficiency, many distributed systems that deploy machine learning solutions in general and RL solutions have started to integrate fog computing in the systems' structure.

So, it can be very promising to integrate fog computing with RL applications deployment in EV charging distributed systems. The EVs will act as the edge devices, which is the very lower layer in the hierarchy. Furthermore, the chargers (EVCS) will act as the middle layer and it can be divided into sub-layers depending on the needed architecture. And at the top, there will be the centralized EV charging service operator centers that are usually connected to the cloud. All of these identities will be considered learning agents in the environment. MARL solutions which are promising solutions will now have a well-structured system for information sharing between the multiple learning agents in the environment, which is going to lead a very high probability of converging to an optimal policy.

A few papers actually have shed light upon the tremendous gain that can be achieved with the enhanced computational structure of the EV charging-based systems. The very recent study [119] has shown some results of considerably reducing the EV charging waiting time and effectively managing the EV charging scheduling using the fog computational structure. Also, another recent study [120] has shown that using the fog hierarchy in the EV to Grid systems can have very beneficial applications in the future. So, this integration between the fog computing and the RL may have the opportunity to shape most of the future applications deployed in the EV charging systems. It is expected that in the few coming years, this integration will dominate the literature of EV charging systems.

## VI. CONCLUSION

The introduction of PEVs in distributed energy systems can address the challenges of managing intermittent renewable energy resources. However, uncertainties are emerged from introducing new flexibilities to the system. Traditional approaches such as optimization techniques have been used to solve the issues of uncertainties in the PEV. However, it is not possible to mathematically model the uncertainties using optimization techniques. RL approaches have proven their ability to interact with the environment of the distributed energy systems to learn the optimal solution to the EV management problem. Solving with RL showed the opportunity of applying the approaches in real-life. In this paper, we performed a comparative analysis of the RL methods used in the EV management system.

Different aspects of the algorithms used were surveyed and compared. Q-learning and its deep extension DQN proved to be the most useful algorithms to tackle the problem. The integration of DQN with DDPG solved both the dilemma of high dimensionality and discretization. Also, the differences between using single-agent RL and MARL approaches in the context of EV charging systems were discussed in this paper. Single-agent RL solutions deployed in EVCS are very likely to converge to a sub-optimal policy due to the limited environment representation perceived by the single agent. MARL solutions deployed in EVs can converge to optimal policies since actions are taken jointly by multiple agents in the same environment, but they are computationally expensive and more difficult to converge.

Different combinations of state parameters were introduced and related to the objectives of the papers surveyed.

The formulation of the states, rewards, and actions used to represent the environment is a very crucial step in the RL solution. The input features are chosen based on the objective of the solution and the particular part of the system to achieve the intended goal. The surveyed research showed that solving different parts of the problem may lead to different solutions. Most of the papers assigned constraints and included them in the cost function of the algorithm to produce practical models that do not violate system physical limitations.

The RL methods proved to perform better than the traditional centralized methods, requiring less computational power and providing faster solutions. However, more research is needed to apply RL techniques in real-life distributed energy systems, as most of the results of the algorithms were obtained through testing against historical data or sophisticated simulations. Nevertheless, the integration of RL techniques and other advanced technologies such as the fog computational architecture shows promise in providing more reliable, and realistic models.

## REFERENCES

[1] M. Yilmaz and P. T. Krein, "Review of the impact of vehicle-to-grid technologies on distribution systems and utility interfaces," *IEEE Trans. Power Electron.*, vol. 28, no. 12, pp. 5673–5689, Dec. 2013, doi: 10.1109/TPEL.2012.2227500.

[2] X. Qi, G. Wu, K. Boriboonsomsin, M. J. Barth, and J. Gonder, "Data-driven reinforcement learning–based real-time energy management system for plug-in hybrid electric vehicles," *Transp. Res. Rec.*, vol. 2572, no. 1, pp. 1–8, Jan. 2016, doi: 10.3141/2572-01.

[3] K. Clement-Nyns, E. Haesen, and J. Driesen, "The impact of charging plug-in hybrid electric vehicles on a residential distribution grid," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 371–380, Feb. 2010, doi: 10.1109/TPWRS.2009.2036481.

[4] L. Pieltain Fernandez, T. Gomez San Roman, R. Cossent, C. Mateo Domingo, and P. Frias, "Assessment of the impact of plug-in electric vehicles on distribution networks," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 206–213, Feb. 2011, doi: 10.1109/TPWRS.2010.2049133.

[5] S. Shao, M. Pipattanasomporn, and S. Rahman, "Challenges of PHEV penetration to the residential distribution network," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2009, pp. 1–8, doi: 10.1109/PES.2009.5275806.

[6] M. Moghimi, D. Leskarac, N. Nadian, S. Stegen, and J. Lu, "Impact of PEV behavior on peak demand reduction in a commercial microgrid," in *Proc. Australas. Universities Power Eng. Conf. (AUPEC)*, Sep. 2016, pp. 1–5, doi: 10.1109/AUPEC.2016.7749380.

[7] B. W. Zhou, T. Littler, and H. F. Wang, "The impact of vehicle-to-grid on electric power systems: A review," in *Proc. 2nd IET Renew. Power Gener. Conf. (RPG)*, 2013, p. 623, doi: 10.1049/cp.2013.1783.

[8] L. Wang and B. Chen, "Model-based analysis of V2G impact on battery degradation," SAE Tech. Papers 2017-01-1699, Mar. 2017, doi: 10.4271/2017-01-1699.

[9] J. J. A. Saldanha, E. M. dos Santos, A. P. C. de Mello, and D. P. Bernardon, "Control strategies for smart charging and discharging of plug-in electric vehicles," in *Smart Cities Technologies*, Rijeka, Croatia: InTech, 2016.

[10] P.-Y. Kong and G. K. Karagiannidis, "Charging schemes for plug-in hybrid electric vehicles in smart grid: A survey," *IEEE Access*, vol. 4, pp. 6846–6875, 2016, doi: 10.1109/ACCESS.2016.2614689.

[11] A. S. Al-Ogaili, T. J. Tengku Hashim, N. A. Rahmat, A. K. Ramasamy, M. B. Marsadek, M. Faisal, and M. A. Hannan, "Review on scheduling, clustering, and forecasting strategies for controlling electric vehicle charging: Challenges and recommendations," *IEEE Access*, vol. 7, pp. 128353–128371, 2019, doi: 10.1109/ACCESS.2019.2939595.

[12] C. Cao, L. Wang, and B. Chen, "Mitigation of the impact of high plug-in electric vehicle penetration on residential distribution grid using smart charging strategies," *Energies*, vol. 9, no. 12, p. 1024, Dec. 2016, doi: 10.3390/en9121024.

[13] A. Chakrabortty and M. D. Ilić, *Control and Optimization Methods for Electric Smart Grids*. New York, NY, USA: Springer, 2011.

[14] T. Bevis, B. Hacker, C. S. Edrington, and S. Azongha, "A review of PHEV grid impacts," in *Proc. 41st North Amer. Power Symp.*, Oct. 2009, pp. 1–6, doi: 10.1109/NAPS.2009.5483995.

[15] C. Zhang, Q. Huang, J. Tian, L. Chen, Y. Cao, and R. Zhang, "Smart grid facing the new challenge: The management of electric vehicle charging loads," *Energy Procedia*, vol. 12, pp. 98–103, 2011, doi: 10.1016/j.egypro.2011.10.014.

[16] N. Shaukat, B. Khan, S. M. Ali, C. A. Mehmood, J. Khan, U. Farid, M. Majid, S. M. Anwar, M. Jawad, and Z. Ullah, "A survey on electric vehicle transportation within smart grid system," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1329–1349, Jan. 2018, doi: 10.1016/j.rser.2017.05.092.

[17] S. Deilami, A. S. Masoum, P. S. Moses, and M. A. S. Masoum, "Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 456–467, Sep. 2011, doi: 10.1109/TSG.2011.2159816.

[18] A. S. Masoum, S. Deilami, P. S. Moses, M. A. S. Masoum, and A. Abu-Siada, "Smart load management of plug-in electric vehicles in distribution and residential networks with charging stations for peak shaving and loss minimisation considering voltage regulation," *IET Gener., Transmiss. Distrib.*, vol. 5, no. 8, pp. 877–888, Aug. 2011, doi: 10.1049/iet-gtd.2010.0574.

[19] W. Tang and Y. J. Zhang, "A model predictive control approach for low-complexity electric vehicle charging scheduling: Optimality and scalability," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1050–1063, Mar. 2017, doi: 10.1109/TPWRS.2016.2585202.

[20] L. Gan, U. Topcu, and S. H. Low, "Optimal decentralized protocol for electric vehicle charging," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 940–951, May 2013, doi: 10.1109/TPWRS.2012.2210288.

[21] S. Sojoudi and S. H. Low, "Optimal charging of plug-in hybrid electric vehicles in smart grids," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2011, pp. 1–6, doi: 10.1109/PES.2011.6039236.

[22] E. Sortomme, M. M. Hindi, S. D. J. MacPherson, and S. S. Venkata, "Coordinated charging of plug-in hybrid electric vehicles to minimize distribution system losses," *IEEE Trans. Smart Grid*, vol. 2, no. 1, pp. 198–205, Mar. 2011, doi: 10.1109/TSG.2010.2090913.

[23] L. Rao, X. Liu, L. Xie, and Z. Pang, "Hedging against uncertainty: A tale of Internet data center operations under smart grid environment," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 555–563, Sep. 2011, doi: 10.1109/TSG.2011.2159523.

[24] T. Jiang, G. Putrus, Z. Gao, M. Conti, S. McDonald, and G. Lacey, "Development of a decentralized smart charge controller for electric vehicles," *Int. J. Electr. Power Energy Syst.*, vol. 61, pp. 355–370, Oct. 2014, doi: 10.1016/j.ijepes.2014.03.023.

[25] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, "Machine learning approaches for EV charging behavior: A review," *IEEE Access*, vol. 8, pp. 168980–168993, 2020, doi: 10.1109/ACCESS.2020.3023388.

[26] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.

[27] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018, doi: 10.17775/cseejpes.2018.00520.

[28] H. Nguyen, L. Kieu, T. Wen, and C. Cai, "Deep learning methods in transportation domain: A review," *IET Intell. Transp. Syst.*, vol. 12, no. 9, pp. 998–1004, Nov. 2018, doi: 10.1049/iet-its.2018.0064.

[29] Z.-J.-M. Shen, B. Feng, C. Mao, and L. Ran, "Optimization models for electric vehicle service operations: A literature review," *Transp. Res. B, Methodol.*, vol. 128, pp. 462–477, Oct. 2019, doi: 10.1016/j.trb.2019.08.006.

[30] A. Ahmad, Z. A. Khan, M. Saad Alam, and S. Khateeb, "A review of the electric vehicle charging techniques, standards, progression and evolution of EV technologies in Germany," *Smart Sci.*, vol. 6, no. 1, pp. 36–53, Jan. 2018, doi: 10.1080/23080477.2017.1420132.

[31] Y. Zhang, X. Liu, T. Zhang, and Z. Gu, "Review of the electric vehicle charging station location problem," in *Dependability in Sensor, Cloud, and Big Data Systems and Applications*. Singapore: Singapore, 2019, pp. 435–445.

[32] J. García-Villalobos, I. Zamora, J. I. San Martín, F. J. Asensio, and V. Aperribay, "Plug-in electric vehicles in electric distribution networks: A review of smart charging approaches," *Renew. Sustain. Energy Rev.*, vol. 38, pp. 717–731, Oct. 2014, doi: 10.1016/j.rser.2014.07.040.

[33] O. Veneri, *Technologies and Applications for Smart Charging of Electric and Plug-in Hybrid Vehicles*. New York, NY, USA: Springer, 2016.

[34] S. K. Rathor and D. Saxena, "Energy management system for smart grid: An overview and key issues," *Int. J. Energy Res.*, vol. 44, no. 6, pp. 4067–4109, May 2020, doi: 10.1002/er.4883.

[35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[36] J. D. Johnson, J. Li, and Z. Chen, "Reinforcement learning: An introduction: R.S. Sutton, A.G. Barto, MIT Press, Cambridge, MA 1998, 322 pp. ISBN 0-262-19398-1," *Neurocomputing*, vol. 35, nos. 1–4, pp. 205–206, 2000, doi: 10.1016/S0925-2312(00)00324-6.

[37] R. S. Sutton, "Dyna, an integrated architecture for learning, planning, and reacting," *ACM SIGART Bull.*, vol. 2, no. 4, pp. 160–163, Jul. 1991, doi: 10.1145/122344.122377.

[38] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou, "Information theoretic MPC for model-based reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1714–1721, doi: 10.1109/ICRA.2017.7989202.

[39] R. F. Baumeister and M. R. Leary, "Writing narrative literature reviews," *Rev. Gen. Psychol.*, vol. 1, no. 3, pp. 311–320, Sep. 1997, doi: 10.1037/1089-2680.1.3.311.

[40] Y. He, B. Venkatesh, and L. Guan, "Optimal scheduling for charging and discharging of electric vehicles," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1095–1105, Sep. 2012, doi: 10.1109/TSG.2011.2173507.

[41] R.-C. Leou, C.-L. Su, and C.-N. Lu, "Stochastic analyses of electric vehicle charging impacts on distribution network," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1055–1063, May 2014, doi: 10.1109/TPWRS.2013.2291556.

[42] L. Rao and J. Yao, "SmartCar: Smart charging and driving control for electric vehicles in the smart grid," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 2709–2714, doi: 10.1109/GLOCOM.2014.7037217.

[43] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Eng. Pract.*, vol. 11, no. 7, pp. 733–764, Jul. 2003, doi: 10.1016/S0967-0661(02)00186-7.

[44] L. Gan, A. Wierman, U. Topcu, N. Chen, and S. H. Low, "Real-time deferrable load control: Handling the uncertainties of renewable generation," in *Proc. 4th Int. Conf. Future Energy Syst.*, 2013, pp. 113–124.

[45] N. Chen, L. Gan, S. H. Low, and A. Wierman, "Distributional analysis for model predictive deferrable load control," in *Proc. 53rd IEEE Conf. Decis. Control*, Dec. 2014, pp. 6433–6438, doi: 10.1109/CDC.2014.7040398.

[46] S. Bansal, M. N. Zeilinger, and C. J. Tomlin, "Plug-and-play model predictive control for electric vehicle charging and voltage control in smart grids," in *Proc. 53rd IEEE Conf. Decis. Control*, Dec. 2014, pp. 5894–5900, doi: 10.1109/CDC.2014.7040312.

[47] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, Jul. 1966, doi: 10.1126/science.153.3731.34.

[48] 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 Gener., Transmiss. Distrib.*, vol. 8, no. 6, pp. 1007–1016, 2014, doi: 10.1049/iet-gtd.2013.0624.

[49] J. Kang, S. J. Duncan, and D. N. Mavris, "Real-time scheduling techniques for electric vehicle charging in support of frequency regulation," *Procedia Comput. Sci.*, vol. 16, pp. 767–775, Jan. 2013, doi: 10.1016/j.procs.2013.01.080.

[50] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[51] M. J. S. Daniel P. Heyman, *Stochastic Models in Operations Research: Stochastic Optimization*, vol. 2, Mineola, NY, USA: Dover, 2004.

[52] F. J. Soares, D. Rua, C. Gouveia, B. D. Tavares, A. M. Coelho, and J. A. P. Lopes, "Electric vehicles charging: Management and control strategies," *IEEE Veh. Technol. Mag.*, vol. 13, no. 1, pp. 130–139, Mar. 2018, doi: 10.1109/MVT.2017.2781538.

[53] C. D. Korkas, S. Baldi, S. Yuan, and E. B. Kosmatopoulos, "An adaptive learning-based approach for nearly optimal dynamic charging of electric vehicle fleets," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 7, pp. 2066–2075, Jul. 2018, doi: 10.1109/TITS.2017.2737477.

[54] K. Valogianni, W. Ketter, and J. Collins, "Smart charging of electric vehicles using reinforcement learning," in *Proc. AAAI Workshop-Tech. Rep.* Bellevue, WA, USA: Trading Agent Des. Anal., vol. WS-13-15, 2013, pp. 41–48.

[55] J. Liu, H. Guo, J. Xiong, N. Kato, J. Zhang, and Y. Zhang, "Smart and resilient EV charging in SDN-enhanced vehicular edge computing networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 1, pp. 217–228, Jan. 2020, doi: 10.1109/JSAC.2019.2951966.

[56] A. Chiş, J. Lundén, and V. Koivunen, "Reinforcement learning-based plug-in electric vehicle charging with forecasted price," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3674–3684, May 2017, doi: 10.1109/TVT.2016.2603536.

[57] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, and G. Deconinck, "Reinforcement learning of heuristic EV fleet charging in a day-ahead electricity market," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1795–1805, Jul. 2015, doi: 10.1109/TSG.2015.2393059.

[58] S. Wang, S. Bi, and Y. A. Zhang, "Reinforcement learning for real-time pricing and scheduling control in EV charging stations," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 849–859, Feb. 2021, doi: 10.1109/tii.2019.2950809.

[59] T. Ding, Z. Zeng, J. Bai, B. Qin, Y. Yang, and M. Shahidehpour, ‘‘Optimal electric vehicle charging strategy with Markov decision process and reinforcement learning technique,’’ *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5811–5823, Sep. 2020, doi: 10.1109/tia.2020.2990096.

[60] W. Shi and V. W. S. Wong, ‘‘Real-time vehicle-to-grid control algorithm under price uncertainty,’’ in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2011, pp. 261–266, doi: 10.1109/SmartGridComm.2011.6102330.

[61] S. Dimitrov and R. Lguensat, ‘‘Reinforcement learning based algorithm for the maximization of EV charging station revenue,’’ in *Proc. Int. Conf. Math. Comput. Sci. Ind.*, Sep. 2014, pp. 235–239, doi: 10.1109/MCSI.2014.54.

[62] A. Chiş, J. Lundén, and V. Koivunen, ‘‘Scheduling of plug-in electric vehicle battery charging with price prediction,’’ in *Proc. IEEE PES ISGT Eur.*, Oct. 2013, pp. 1–5, doi: 10.1109/ISGTEurope.2013.6695263.

[63] J. Lee, E. Lee, and J. Kim, ‘‘Electric vehicle charging and discharging algorithm based on reinforcement learning with data-driven approach in dynamic pricing scheme,’’ *Energies*, vol. 13, no. 8, p. 1950, Apr. 2020, doi: 10.3390/en13081950.

[64] H. Li, Z. Wan, and H. He, ‘‘Constrained EV charging scheduling based on safe deep reinforcement learning,’’ *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, May 2020, doi: 10.1109/tsg.2019.2955437.

[65] 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 Trans. Smart Grid*, vol. 11, no. 2, pp. 1714–1723, Mar. 2020, doi: 10.1109/TSG.2019.2942593.

[66] J. Shi, Y. Gao, W. Wang, N. Yu, and P. A. Ioannou, ‘‘Operating electric vehicle fleet for ride-hailing services with reinforcement learning,’’ *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4822–4834, Nov. 2020, doi: 10.1109/TITS.2019.2947408.

[67] F. Zhang, Q. Yang, and D. An, ‘‘CDDPG: A deep-reinforcement-learning-based approach for electric vehicle charging control,’’ *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3075–3087, Mar. 2021, doi: 10.1109/jiot.2020.3015204.

[68] S. Najafi, M. Shafie-khah, P. Siano, W. Wei, and J. P. S. Catalão, ‘‘Reinforcement learning method for plug-in electric vehicle bidding,’’ *IET Smart Grid*, vol. 2, no. 4, pp. 529–536, 2019, doi: 10.1049/iet-stg.2018.0297.

[69] M. Shin, D.-H. Choi, and J. Kim, ‘‘Cooperative management for PV/ESS-enabled electric vehicle charging stations: A multiagent deep reinforcement learning approach,’’ *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 3493–3503, May 2020, doi: 10.1109/TII.2019.2944183.

[70] F. L. D. Silva, C. E. H. Nishida, D. M. Roijers, and A. H. R. Costa, ‘‘Coordination of electric vehicle charging through multiagent reinforcement learning,’’ *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2347–2356, May 2020, doi: 10.1109/TSG.2019.2952331.

[71] I. Dusparic, C. Harris, A. Marinescu, V. Cahill, and S. Clarke, ‘‘Multi-agent residential demand response based on load forecasting,’’ in *Proc. 1st IEEE Conf. Technol. Sustainability (SusTech)*, Aug. 2013, pp. 90–96, doi: 10.1109/SusTech.2013.6617303.

[72] I. Dusparic, A. Taylor, A. Marinescu, V. Cahill, and S. Clarke, ‘‘Maximizing renewable energy use with decentralized residential demand response,’’ in *Proc. IEEE 1st Int. Smart Cities Conf. (ISC2)*, Oct. 2015, pp. 1–6, doi: 10.1109/ISC2.2015.7366212.

[73] C. X. Jiang, Z. X. Jing, X. R. Cui, T. Y. Ji, and Q. H. Wu, ‘‘Multiple agents and reinforcement learning for modelling charging loads of electric taxis,’’ *Appl. Energy*, vol. 222, pp. 158–168, Jul. 2018, doi: 10.1016/j.apenergy.2018.03.164.

[74] D. Dauer, C. M. Flath, P. Strohle, and C. Weinhardt, ‘‘Market-based EV charging coordination,’’ in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, Nov. 2013, pp. 102–107, doi: 10.1109/WI-IAT.2013.97.

[75] A. Di Giorgio, F. Liberati, and A. Pietrabissa, ‘‘On-board stochastic control of electric vehicle recharging,’’ in *Proc. 52nd IEEE Conf. Decis. Control*, Dec. 2013, pp. 5710–5715, doi: 10.1109/CDC.2013.6760789.

[76] B. J. Claessens, S. Vandael, F. Ruelens, K. De Craemer, and B. Beusen, ‘‘Peak shaving of a heterogeneous cluster of residential flexibility carriers using reinforcement learning,’’ in *Proc. IEEE PES ISGT Eur.*, Oct. 2013, pp. 1–5, doi: 10.1109/ISGTEurope.2013.6695254.

[77] M. G. Vayá, L. B. Roselló, and G. Andersson, ‘‘Optimal bidding of plug-in electric vehicles in a market-based control setup,’’ in *Proc. Power Syst. Comput. Conf.*, Aug. 2014, pp. 1–8, doi: 10.1109/PSCC.2014.7038108.

[78] A. Taylor, I. Dusparic, E. Galvan-Lopez, S. Clarke, and V. Cahill, ‘‘Accelerating learning in multi-objective systems through transfer learning,’’ in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 2298–2305, doi: 10.1109/IJCNN.2014.6889438.

[79] A. Chis, J. Lunden, and V. Koivunen, ‘‘Optimization of plug-in electric vehicle charging with forecasted price,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2086–2089, doi: 10.1109/ICASSP.2015.7178338.

[80] Z. Wen, D. O'Neill, and H. Maei, ‘‘Optimal demand response using device-based reinforcement learning,’’ *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2312–2324, Sep. 2015, doi: 10.1109/TSG.2015.2396993.

[81] A. I. Arif, M. Babar, T. P. I. Ahamed, E. A. Al-Ammar, P. H. Nguyen, I. G. R. Kamphuis, and N. H. Malik, ‘‘Online scheduling of plug-in vehicles in dynamic pricing schemes,’’ *Sustain. Energy, Grids Netw.*, vol. 7, pp. 25–36, Sep. 2016, doi: 10.1016/j.segan.2016.05.001.

[82] A. Marinescu, I. Dusparic, and S. Clarke, ‘‘Prediction-based multi-agent reinforcement learning in inherently non-stationary environments,’’ *ACM Trans. Auto. Adapt. Syst.*, vol. 12, no. 2, pp. 1–23, May 2017, doi: 10.1145/3070861.

[83] H. Ko, S. Pack, and V. C. M. Leung, ‘‘Mobility-aware vehicle-to-grid control algorithm in microgrids,’’ *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 7, pp. 2165–2174, Jul. 2018, doi: 10.1109/TITS.2018.2816935.

[84] Z. Wan, H. Li, H. He, and D. Prokhorov, ‘‘Model-free real-time EV charging scheduling based on deep reinforcement learning,’’ *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019, doi: 10.1109/TSG.2018.2879572.

[85] N. Sadeghianpourhamami, J. Deleu, and C. Develder, ‘‘Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning,’’ *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 203–214, Jan. 2020, doi: 10.1109/TSG.2019.2920320.

[86] S. Kim and H. Lim, ‘‘Reinforcement learning based energy management algorithm for smart energy buildings,’’ *Energies*, vol. 11, no. 8, p. 2010, Aug. 2018, doi: 10.3390/en11082010.

[87] S. Lee and D.-H. Choi, ‘‘Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic system, energy storage system, and home appliances,’’ *Sensors*, vol. 19, no. 18, p. 3937, Sep. 2019, doi: 10.3390/s19183937.

[88] C. Zhang, S. R. Kuppannagari, C. Xiong, R. Kannan, and V. K. Prasanna, ‘‘A cooperative multi-agent deep reinforcement learning framework for real-time residential load scheduling,’’ in *Proc. Int. Conf. Internet Things Design Implement.*, Apr. 2019, pp. 59–69, doi: 10.1145/3302505.3310069.

[89] K. L. Lopez, ‘‘A machine learning approach for the smart charging of electric vehicles,’’ Univ. Laval, Quebec City, QC, Canada, 2019. Accessed: Dec. 12, 2020. [Online]. Available: http://hdl.handle.net/20.500.11794/34741

[90] R. Glatt, ‘‘Enabling optimized charging of electric vehicles in mobility services,’’ Lawrence Livermore Nat. Lab., Livermore, CA, USA, Tech. Rep. LLNL-TR-800857, Dec. 2019, doi: 10.2172/1581881.

[91] J. Sun, Y. Zheng, J. Hao, Z. Meng, and Y. Liu, ‘‘Continuous multiagent control using collective behavior entropy for large-scale home energy management,’’ in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, vol. 34, no. 1, pp. 922–929.

[92] X. Cui, Y. You, Z. Jing, and J. Zhu, ‘‘Electric taxi charging guidance strategy based on electric vehicle agent-based simulation,’’ in *Proc. DEStech Trans. Comput. Sci. Eng. (MSO)*, 2018, pp. 105–109, doi: 10.12783/dtcse/mso2018/20465.

[93] L. Buechler. (2018). *Aggregated Electric Vehicle Charging Control for Power Grid Ancillary Service Provision*. [Online]. Available: https://web.stanford.edu/class/aa228/reports/2018/final10.pdf

[94] K. Valogianni, W. Ketter, J. Collins, and D. Zhdanov, ‘‘Effective management of electric vehicle storage using smart charging,’’ *Proc. Nat. AAAI Conf. Artif. Intell.*, vol. 1, 2014, pp. 472–478.

[95] Y. Zhang, Z. Zhang, Q. Yang, D. An, D. Li, and C. Li, ‘‘EV charging bidding by multi-DQN reinforcement learning in electricity auction market,’’ *Neurocomputing*, vol. 397, pp. 404–414, Jul. 2020, doi: 10.1016/j.neucom.2019.08.106.

[96] Y. Yang, J. Hao, Y. Zheng, X. Hao, and B. Fu, ‘‘Large-scale home energy management using entropy-based collective multiagent reinforcement learning framework,’’ in *Proc. 18th Int. Conf. Auton. Agents MultiAgent Syst. (AAMAS)*, vol. 4, 2019, pp. 2285–2287.

[97] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, ‘‘A deep reinforcement learning method for pricing electric vehicles with discrete charging levels,’’ *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5901–5912, Sep. 2020, doi: 10.1109/TIA.2020.2984614.

[98] N. Mhaisen, N. Fetais, and A. Massoud, ''Real-time scheduling for electric vehicles charging/discharging using reinforcement learning,'' in *Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIoT)*, Feb. 2020, pp. 1–6, doi: 10.1109/ICIoT48696.2020.9089471.

[99] V. Moghaddam, A. Yazdani, H. Wang, D. Parlevliet, and F. Shahnia, ''An online reinforcement learning approach for dynamic pricing of electric vehicle charging stations,'' *IEEE Access*, vol. 8, pp. 130305–130313, 2020, doi: 10.1109/ACCESS.2020.3009419.

[100] X. Fang, J. Wang, G. Song, Y. Han, Q. Zhao, and Z. Cao, ''Multi-agent reinforcement learning approach for residential microgrid energy scheduling,'' *Energies*, vol. 13, no. 1, p. 123, Dec. 2019, doi: 10.3390/en13010123.

[101] X. S. Zhang, T. Yu, Z. N. Pan, B. Yang, and T. Bao, ''Lifelong learning for complementary generation control of interconnected power grids with high-penetration renewables and EVs,'' *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 4097–4110, Jul. 2018, doi: 10.1109/TPWRS.2017.2767318.

[102] T. Qian, C. Shao, X. Li, X. Wang, and M. Shahidehpour, ''Enhanced coordinated operations of electric power and transportation networks via EV charging services,'' *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3019–3030, Jul. 2020, doi: 10.1109/TSG.2020.2969650.

[103] M. Lahariya, N. Sadeghianpourhamami, and C. Develder, ''Reduced state space and cost function in reinforcement learning for demand response control of multiple EV charging stations,'' in *Proc. 6th ACM Int. Conf. Syst. Energy-Efficient Buildings, Cities, Transp.*, Nov. 2019, pp. 344–345, doi: 10.1145/3360322.3360992.

[104] Q. Dang, D. Wu, and B. Boulet, ''A Q-learning based charging scheduling scheme for electric vehicles,'' in *Proc. IEEE Transp. Electrific. Conf. Expo (ITEC)*, Jun. 2019, pp. 1–5, doi: 10.1109/ITEC.2019.8790603.

[105] C. Böhringer, ''The synthesis of bottom-up and top-down in energy policy modeling,'' *Energy Econ.*, vol. 20, no. 3, pp. 233–248, 1998, doi: 10.1016/S0140-9883(97)00015-7.

[106] D. Ernst, P. Geurts, and L Wehenkel, ''Tree-based batch mode reinforcement learning,'' *J. Mach. Learn. Res.*, vol. 6, pp. 503–556, Jun. 2005.

[107] A. M. Andrew, ''Reinforcement learning: An introduction by Richard S. Sutton and Andrew G. Barto, Adaptive Computation and Machine Learning series, MIT Press (Bradford Book), Cambridge, Mass., 1998, xviii + 322 pp, ISBN 0-262-19398-1, (hardback, $31.95).,'' *Robotica*, vol. 17, no. 2, pp. 229–235, Mar. 1999, doi: 10.1017/s0263574799211174.

[108] J. Van Roy, N. Leemput, S. De Breucker, F. Geth, P. Tant, and J. Driesen. (2011). *An Availability Analysis and Energy Consumption Model for a Flemish Fleet of Electric Vehicles*. Accessed: Oct. 14, 2020. [Online]. Available: https://lirias.kuleuven.be/1730977

[109] M. Kefayati and R. Baldick, ''Energy delivery transaction pricing for flexible electrical loads,'' in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2011, pp. 363–368, doi: 10.1109/SmartGridComm.2011.6102348.

[110] E. L. Karfopoulos and N. D. Hatziargyriou, ''A multi-agent system for controlled charging of a large population of electric vehicles,'' *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1196–1204, May 2013, doi: 10.1109/TPWRS.2012.2211624.

[111] E. L. Karfopoulos, P. Papadopoulos, S. Skarvelis-Kazakos, I. Grau, L. M. Cipcigan, N. Hatziargyriou, and N. Jenkins, ''Introducing electric vehicles in the microgrids concept,'' in *Proc. 16th Int. Conf. Intell. Syst. Appl. Power Syst.*, Sep. 2011, pp. 1–6, doi: 10.1109/ISAP.2011.6082212.

[112] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. London, U.K.: Academic, 1999.

[113] *Overview (JADE V4)*. Accessed: Oct. 26, 2020. [Online]. Available: https://jade.tilab.com/doc/api/index.html

[114] Q. Huang, Q.-S. Jia, and X. Guan, ''Robust scheduling of EV charging load with uncertain wind power integration,'' *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1043–1054, Mar. 2018, doi: 10.1109/TSG.2016.2574799.

[115] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra, ''Residential demand response using reinforcement learning,'' in *Proc. 1st IEEE Int. Conf. Smart Grid Commun.*, Oct. 2010, pp. 409–414, doi: 10.1109/smartgrid.2010.5622078.

[116] Y. Ge, F. Zhu, X. Ling, and Q. Liu, ''Safe Q-learning method based on constrained Markov decision processes,'' *IEEE Access*, vol. 7, pp. 165007–165017, 2019, doi: 10.1109/ACCESS.2019.2952651.

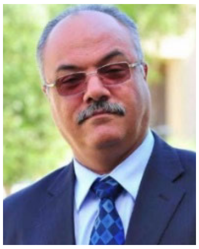
[117] *SCIP*. Accessed: Oct. 23, 2020. [Online]. Available: https://www.scipopt.org/

[118] H. Song, C.-C. Liu, J. Lawarree, and R. W. Dahlgren, ''Optimal electricity supply bidding by Markov decision process,'' *IEEE Trans. Power Syst.*, vol. 15, no. 2, pp. 618–624, May 2000, doi: 10.1109/59.867150.

[119] D. A. Chekired, L. Khoukhi, and H. T. Mouftah, ''Fog-computing-based energy storage in smart grid: A cut-off priority queuing model for plug-in electrified vehicle charging,'' *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 3470–3482, May 2020, doi: 10.1109/TII.2019.2940410.

[120] N. Kumar, T. Dhand, A. Jindal, G. S. Aujla, H. Cao, and L. Yang, ''An edge-fog computing framework for cloud of things in vehicle to grid environment,'' in *Proc. IEEE 21st Int. Symp. 'A World Wireless, Mobile Multimedia Netw.' (WoWMoM)*, Aug. 2020, pp. 354–359, doi: 10.1109/WoWMoM49955.2020.00066.

**HEBA M. ABDULLAH** (Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Kuwait University, in July 2007 and May 2018, respectively. She is currently pursuing the Ph.D. degree with Qatar University. She has gained knowledge from working with international consultation companies whose clients are Kuwait Oil Company (KOC), Kuwait National Petroleum Company (KNPC), and the Kuwait Institute for Scientific Research (KISR). She also works as a Consultant in the areas of power systems, renewable energy, and sustainable solutions. She is also a Researcher with Kuwait University, Qatar University, the University of Exeter, PAAET, and the Founder of ReHub United Consultation and the Research Group.



**ADEL GASTLI** (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the National School of Engineers of Tunis, Tunisia, in 1985, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Nagoya Institute of Technology, Japan, in March 1990 and March 1993, respectively. From September 1985 to September 1987, he has worked with the National Institute for Standards and Intellectual Property, Tunisia. He has worked with Mitsubishi Electric Corporation, Japan, from April 1993 to July 1995. He joined the Electrical and Computer Engineering Department, Sultan Qaboos University, Oman, in August 1995. He has served as the Head of the Department from September 2001 to August 2003 and September 2007 to August 2009. He was appointed as the Director of the Sultan Qaboos University Quality Assurance Office from February 2010 to January 2013. In February 2013, he joined the Electrical Engineering Department, Qatar University, as a Professor and the Kahramaa-Siemens Chair in energy efficiency. From August 2013 to September 2015, he was appointed as the College of Engineering Associate Dean for Academic Affairs. His current research interests include energy efficiency, renewable energy, electric vehicles, and smart grid.

**LAZHAR BEN-BRAHIM** (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the National School of Engineers of Tunis, Tunisia, in 1985 and 1987, respectively, and the Ph.D. degree in electrical engineering from Yokohama National University, Yokohama, Japan, in 1991. From 1991 to 1997, he was with Toshiba Corporation, where he was engaged in research and development of power electronics and motor drive systems. In September 1997, he joined the Industrial Technology Department, College of Technology, Qatar University. He has served as the Head of the Industrial Technology Department, from 1998 to 2005. In September 2005, he joined the Electrical Engineering Department, Qatar University. He was also the Industrial Electronics Chair for RasGas Company and the Head of the Electrical Engineering Department. He invented several new techniques for use in motor drives, power electronics, sensors, and related areas. These inventions are registered in more than 12 international patents. His current research interests include power electronics, renewable energy, electric vehicles, electric drives, and sensor and instrumentation. He is a member of IEE Japan. He is an Associate Editor of the *Journal of Electrical Engineering* (Springer) and an Editor of the *Electronics* journal (MDPI).