energies

MDPI

Review

Hosting Capacity Assessment Strategies and Reinforcement Learning Methods for Coordinated Voltage Control in Electricity Distribution Networks: A Review

Jude Suchithra [1], Duane Robinson [1,*] and Amin Rajabi [2]

[1] Australian Power Quality and Reliability Centre, University of Wollongong, Wollongong 2522, Australia
[2] DIgSILENT Pacific, Sydney 2000, Australia
* Correspondence: duane@uow.edu.au

Abstract: Increasing connection rates of rooftop photovoltaic (PV) systems to electricity distribution networks has become a major concern for the distribution network service providers (DNSPs) due to the inability of existing network infrastructure to accommodate high levels of PV penetration while maintaining voltage regulation and other operational requirements. The solution to this dilemma is to undertake a hosting capacity (HC) study to identify the maximum penetration limit of rooftop PV generation and take necessary actions to enhance the HC of the network. This paper presents a comprehensive review of two topics: HC assessment strategies and reinforcement learning (RL)-based coordinated voltage control schemes. In this paper, the RL-based coordinated voltage control schemes are identified as a means to enhance the HC of electricity distribution networks. RL-based algorithms have been widely used in many power system applications in recent years due to their precise, efficient and model-free decision-making capabilities. A large portion of this paper is dedicated to reviewing RL concepts and recently published literature on RL-based coordinated voltage control schemes. A non-exhaustive classification of RL algorithms for voltage control is presented and key RL parameters for the voltage control problem are identified. Furthermore, critical challenges and risk factors of adopting RL-based methods for coordinated voltage control are discussed.

Keywords: voltage control; hosting capacity; reinforcement learning; artificial neural networks; quasi-static time series; photovoltaic systems; electricity distribution networks

Citation: Suchithra, J.; Robinson, D.; Rajabi, A. Hosting Capacity Assessment Strategies and Reinforcement Learning Methods for Coordinated Voltage Control in Electricity Distribution Networks: A Review. *Energies* **2023**, *16*, 2371. https://doi.org/10.3390/en16052371

Academic Editors: Alfredo Vaccaro and Fabrizio de Caro

Received: 11 January 2023
Revised: 10 February 2023
Accepted: 24 February 2023
Published: 1 March 2023

1. Introduction

Renewable energy generation plays a vital role in decarbonization of the energy sector and its transition to a green and more sustainable system. According to the International Energy Agency [1], the global renewable capacity is estimated to increase by almost 75% between the years 2022 and 2027. Solar generation from rooftop solar PV panels is one of the most attractive modes of renewable energy generation. Solar PV generation capacity is estimated to become the largest installed electricity capacity worldwide by 2027, surpassing natural gas, coal and hydropower [1]. Transitioning the electricity distribution network to deal with the increasing uptake of rooftop PV systems is a challenging task, as high amounts of PV generation may cause adverse technical or operational network impacts.

According to [2], the definition of PV-hosting capacity (PVHC) is the total PV generation capacity that can be accommodated on a given feeder without any adverse impacts on the electricity distribution network. High penetration of PV generation in an electricity distribution network may violate network performance indices for voltage, thermal limits, protection coordination and power quality. Such network operational constraint violations undermine the network reliability and often result in high economic losses for the DNSPs. Assessing the PVHC is key in mitigating such adverse impacts and ensuring the reliability of the electricity distribution network. Due to the lack of advanced planning approaches by DNSPs, the PVHC can often be under- or overestimated. One of the main objectives of this

paper is to review various methodologies in the literature that quantify the PVHC more accurately and in a time-efficient manner.

The violation of voltage constraints is ubiquitously expressed as the limiting factor in nearly all hosting capacity assessment studies. Innovative approaches to regulate the voltage in the electricity distribution network are the obvious solution to enhance the HC and integrate more distributed energy resources (DERs) into the grid [3]. Voltage control schemes such as Volt-VAr and Volt-Watt of smart inverters only utilize local measurement data to regulate the voltage [4]. In contrast, coordinated voltage control schemes unlock the full potential of the voltage-regulating devices and maximize their performance in regulating the voltage [5]. Such control schemes of network elements have a direct influence on the HC, especially in unbalanced low-voltage (LV) networks.

Due to the recent advancements in the fields of artificial intelligence and machine learning, researchers have focused their attention on reinforcement learning (RL)-based control algorithms to solve various control tasks. RL approaches to solve the coordinated voltage control problem have been given increased attention in recent years due to their ability to make highly accurate and efficient decisions as compared to other voltage control algorithms [6].

Table 1 summarizes the main features of several recently published review works on HC and RL applications for power systems and highlights the main contributions of the current paper. Only a few publications can be found on applications of RL algorithms for quantifying and enhancing the HC of distribution networks, which is identified as a research gap that deserves more attention [7–9].

Table 1. Comparison of prominent features of previous studies and current work.

Features	[10]	[11]	[12]	[13]	[14]	[15]	Current Work
HC assessment methods	✓	✓	x	x	x	x	✓
HC enhancement techniques	x	✓	x	x	x	x	P
An overview of reinforcement learning algorithms	x	x	✓	✓	✓	✓	✓
Coordinated voltage control using reinforcement learning methods	x	x	✓	✓	✓	P	✓
Identification of various challenges in using reinforcement learning methods	x	x	✓	x	✓	P	✓

P = Partially.

This paper discusses the key concepts of HC and RL-based algorithms and provides a guide for researchers pursuing activities on the utilization of RL-based algorithms for HC studies. In summary, this paper offers several contributions compared with previous efforts:

- It categorizes and explains the major RL-based algorithms that are applicable to the power system domain and especially distribution network management;
- It explores a wide range of recent publications on RL-based coordinated voltage control algorithms and highlights the significance of such control methods in enhancing the performance of electricity distribution networks;
- It compares the main features of these algorithms and reviews their advantages and disadvantages;
- It explains the HC assessment methods and the application of RL-based algorithms for enhancing the HC of distribution networks.

The remaining sections of this paper are organized as follows: Section 2 reviews the hosting capacity quantification methods published in recent literature. Section 3 gives a brief introduction to the reinforcement learning concepts. In Section 4, reinforcement learning algorithms for coordinated voltage control in recent literature are classified. Section 5 details multi-agent reinforcement learning methods. Section 6 describes certain parameters of reinforcement learning algorithms that are unique to the voltage control problem. In

Section 7, several risk factors and challenges of reinforcement learning algorithms are identified. Lastly, in Section 8 conclusions are presented.

2. Hosting Capacity Analysis

According to [11], the HC concept originated in 2004 and was used to specify the impacts of DER penetration on distribution systems. HC quantification is an important issue for DNSPs since it determines the new DERs that can be added to the distribution system in the future without any violation of network operating conditions. Before undertaking an HC analysis on a distribution network, acceptable limits for performance indices need to be established. Different performance indices are used for HC quantification studies in various studies, and [11] classifies the main HC performance indices used by most authors into the following four groups:

- Overvoltage problems;
- Overloading and power loss problems;
- Power quality problems;
- Protection problems.

The limits for each of these HC performance indices are based on the standards followed by DNSPs. Ref. [16] provides a comparison of performance limits for over/under voltage, voltage unbalance and harmonics based on different standards. Another example of such performance limits is presented in [17] by the Electric Power Research Institute (EPRI) for HC analysis. In [18], an HC assessment of 50,000 real LV distribution systems is undertaken with over/under voltage, voltage unbalance, conductor thermal capacity and transformer overload as the monitored operational limits. It was identified that overvoltage is the most restrictive performance index in HC quantification and constitutes 61.5% of the incidences of operational limit violations, followed by conductor thermal capacity at 27.7% and voltage unbalance at 9.6%. The results from this study are consistent with the fact that overvoltage and conductor thermal capacity are identified as the main two performance indices considered for the HC analysis in other well-cited literature [10,11].

A number of distinctive methods can be seen in various literature to analyse the HC of electricity distribution networks. Traditional deterministic load-flow-based HC analysis is one of the earliest methods that offers faster computational time but with less accuracy. The most common methods for HC quantification are based on probabilistic load flow (PLF) that captures various uncertainties of the distribution network. Several classifications for HC quantification methods are proposed by various authors in [11,16,19], with a comprehensive review of recent literature and trends in HC assessments. A similar classification to that proposed in [10] is followed in this paper, which classifies the HC quantification methods into three groups: deterministic methods, probabilistic load flow methods and quasi-static time series methods. This classification method captures most of the publications for HC assessment techniques in a non-exhaustive manner that facilitates comparing and contrasting their fundamental concepts.

2.1. Deterministic Methods

Fixed input data are used in deterministic methods to analyse the PV hosting capacity of a distribution network. For example, single input values are used for input data models, such as customer power consumption and PV production, to give single output values [10]. Three rule-based deterministic methods to quantify the PV hosting capacity of a medium voltage (MV) network are presented in [20]. These methods consider a rule-based increment of the utilization factor ratio, which is the assumed PV power to the derived roof potential power of the installed PV on each node of the MV feeder. In [21], a comparison of these three algorithms and an alternative Monte Carlo deterministic method for HC quantification are presented. The deterministic Monte Carlo method is similar in concept to the three methods except that the utilization factor is increased by a random predefined value. The Monte Carlo method is highly accurate compared to the mentioned three methods, but it has a high computational burden and is considered to be slow. These three methods do not

suffer the high computational burden of the Monte Carlo method. Therefore, DNSPs can use these three methods to obtain a quick estimate of the HC of the network.

Traditional deterministic load flow simulation-based methods are used in some studies to quantify the maximum PV hosting capacity of a distribution system [22,23]. Such methods utilize iterative power flow simulations and increase the PV capacity in steps until a violation of operational limits occurs. However, the time-varying behaviour of the network is not modelled in such methods, since fixed values for customer load and PV generation are used in the simulations. Deterministic load flow methods are commonly used in the literature to establish the relationship between PV hosting capacity and other network factors such as customer load, feeder terminals and electric vehicle (EV) loads.

Analytical approaches to deterministic methods are widely used in many works of literature, since they offer a quick determination of HC with less computational burden. Power flow analysis software is not required for such analytic methods, and they can easily be implemented in spreadsheet environments [24]. Another example of an analytical approach for HC quantification is presented in [25] for three different scenarios of distributed generation (DG) placement in the network. Similar to other deterministic methods, the HC is quantified in analytical approaches considering the allowable voltage rise of the feeder and the thermal conductivity of the lines.

2.2. Probabilistic Load Flow Methods

The most preferred approach to model uncertainties due to DER in the literature is by utilizing PLF-based methods. In [19], PLF methods to model uncertainties in the electricity distribution network are classified into numerical approaches and analytical approaches. The Monte Carlo simulation (MCS) is the most common numerical PLF method used by many authors to model the uncertainties in a distribution system. MCS is well known to be a highly accurate stochastic method and it is often used as a benchmark for comparison with other HC quantification methods [26–28]. A drawback of MCS-based methods is their high computational burden with the increasing levels of uncertainty in large distribution networks. A method for estimating the HC of a large distribution network based on an MCS performed only on 1% of randomly selected LV systems is presented in [18], which significantly reduces the computational burden of MCS-based methods. However, analytical approaches to HC quantification are commonly adopted by many authors as a solution to the shortcomings of MCS-based methods.

Analytical approaches perform arithmetic using Probability Density Functions (PDF) of stochastic input variables to solve PLF. Some of the common analytical approaches to PLF use techniques such as the point estimation method (PEM), unscented transformation (UT), convolution and cumulants. PEM is an analytical PLF method based on the statistical data provided by the first few central moments of an uncertain input. PEM computational time is considerably less compared to the MCS, but the accuracy of the solution is sensitive to the complexity of the system [29,30]. In terms of modelling correlated uncertain variables in a power system, the UT method [31] is considered a very appealing approach. The computational burden of the UT method is almost the same as the PEM method, but lower than that of MCS. Convolution methods for PLF perform a convolution operation on PDF of the input variables. This method is most suited for small-scale data systems, while for large systems high amounts of storage and computational power are required [32]. The cumulant-based methods prevent the need to perform the convolution operation when calculating the PDF of a linear combination of several random variables [33].

Most analytical methods are incapable of incorporating system dynamics to perform PLF and the relation between system variables over time is lost. An analytical approach based on a Markov chain Monte Carlo formulation is presented in [34] to realize system dynamics and calculate PLF. Latin Hypercube Sampling with Cholesky Decomposition proposed in [35] is another example of an analytical PLF method that can incorporate system dynamics into the PLF solution.

2.3. Quasi-Static Time Series Methods

The quasi-static time series (QSTS) simulation is defined in the IEEE guide [36] as a sequence of steady-state power flow simulations with time steps of no less than 1 s up to steps of 1 h. In QSTS power flow, discrete controls of power distribution devices can be established that can change their state from one time step to the next. The settling time of a dynamic event and the time period between steady-state power flow solutions determine the accuracy of the QSTS analysis.

An analysis of the QSTS requirements, such as the length of the simulation, input data time resolution and the time step resolution, is presented in [37]. This study identifies that higher-resolution simulations with shorter time periods are more accurate than the simulations with longer time periods with lower resolutions. To produce the most accurate results, QSTS simulation time resolution must be less than the fastest delay of any discrete control device in the distribution system. A QSTS simulation to analyse the impact of PV generation on voltage regulation devices such as on-load tap changers (OLTCs), capacitor banks and voltage regulators is presented in [38]. In this study, high-resolution data for PV output profiles are synthesized from solar irradiance data or the proxy data of similar plants.

Performing QSTS year-long simulations on large distribution systems at high-resolution timescales is proven to be computationally costly. This dilemma has driven researchers to develop analytical techniques to reduce the computational time of QSTS simulations. Ref. [39] presents a method to reduce the feeders in a distribution system to a specified number of buses of interest while retaining its equivalent properties. The proposed method reduces the QSTS simulation time significantly while incorporating multiphase connections, mutual coupling between unbalanced multiphase lines, spatial variations and unbalanced loads/generators into the simulation process.

In [40], vector quantization is utilized to perform fast QSTS year-long simulations to study the impacts of distributed PV generation. Vector quantization identifies similar input data profiles and simulates them once, eliminating the redundant calculation steps and improving the computation time. The traditional algorithm to perform vector quantization is k-means, in which the input data are partitioned into clusters (time steps) and the similarity between clusters is established through Euclidean distance. Down-sampling is another technique to speed up the QSTS simulation by reducing the resolution of the input data profiles. The study presented in [41] undertakes a comparative study between down-sampling and vector quantization methods for shortening the QSTS simulation time. It is identified that vector quantization is the superior technique in terms of accuracy and the simulation time when voltage control is implemented in the simulation.

A scalable and fast QSTS analysis method is proposed in [42] that relies on a linear sensitivity model which exploits the correlation between the real/reactive power injections and the feeder voltage using multiple linear regression. Estimation of control actions is one of the challenges in undertaking QSTS simulations. The proposed sensitivity model is inspired by predictive modelling in machine learning and accurately estimates the control actions required for the entire year-long simulation. The QSTS simulation studies discussed above consider static methods for determining hosting capacity by considering infrequent worst-case snapshots in time. The durations of the violations are not accurately captured by the traditional QSTS simulation methods. Traditional QSTS simulations do not consider instances in which voltage violations are temporarily acceptable, and they often lead to overestimation of the HC. The study presented in [43] proposes a dynamic PV hosting capacity quantification methodology with the formulation of time-aware metrics. The proposed methodology accurately captures the operation of control-based mitigation techniques such as Volt-VAr control that enhances the HC of distribution networks.

2.4. Voltage Control

As highlighted in the preceding sections, voltage constraint violations are recognized by many authors as the primary cause that restricts the HC of electricity distribution

networks. The voltage in traditional distribution systems without DER is based on the radial power flows from the substation to the loads. The introduction of DERs may reduce the performance of the traditional voltage control schemes as the direction of power flows is now reversed. Transitioning into a smart distribution system through active network management of DERs is one of the solutions to resolve the voltage constraint violations and enhance the HC of the network.

Active distributed control schemes require pervasive communication systems to handle the complexity of a smart grid. A classification of such voltage control schemes based on the communication structure is presented in [44] that categorizes voltage control schemes into local control, centralized control, distributed control and decentralized control. Coordinated voltage control schemes exploit the full potential of DERs and other controllable network elements than the local control schemes that only rely on local observations.

The optimal power flow (OPF) problem is at the core of every coordinated voltage control scheme. Some of the commonly used algorithms for coordinated voltage control in literature are listed below:

- Rule-based methods [45–47];
- Analytical methods for OPF [48–50];
- Model-predictive control [51–53];
- Heuristic methods [54–56];
- Reinforcement learning [57–59].

Among these control algorithms, reinforcement learning methods have been adopted in recent years for coordinated voltage control. Recent advancements in artificial neural networks enable deep RL-based control algorithms to produce more desirable results than other control algorithms. The following sections of this paper are dedicated to a brief overview of reinforcement learning concepts and a review of reinforcement learning algorithms proposed in recent literature for coordinated voltage control in electricity distribution networks.

3. An Introduction to Reinforcement Learning

Reinforcement learning is a prominent machine learning paradigm that has been used to solve a variety of control problems in the presence of uncertainty. The application of RL in sustainable energy and electric systems is considered by many authors as a path to revolutionize the traditional power utilization mode and bring more intelligence into power systems. RL is a useful tool that can make (near-) optimal decisions in electricity distribution systems with complex nonlinearity and uncertainty. Some of the key applications of RL in power systems can be seen in the fields of energy management, frequency regulation, voltage control, stability control and congestion management [12,14,15]. However, this paper only reviews applications of RL in the field of coordinated voltage control in electricity distribution systems, since voltage is identified to be the main limiting factor of HC.

3.1. Markov Decision Process

In RL, an agent tries to maximize its cumulative reward by making sequential decisions in an uncertain environment. A Markov decision process (MDP) provides a mathematical framework to formalize sequential decision making in RL. An MDP can be described as a 4-tuple $(S,\ A, P, R)$, where S is the state space $(s \in S)$, A is the action space $(a \in A)$, P is the transition probability of an action a in state s at time t that leads to state s' at time $t+1$ and R is the immediate reward of transitioning to state s' from state s due to action a.

Figure 1 illustrates the MDP where a decision maker called the agent interacts with its environment sequentially. At each time step, given the state of the environment S_t, the agent selects an action A_t. The environment is then transitioned into its new state S_{t+1} and the agent is given an immediate reward R_{t+1} as its consequent action. The state–action pair (s_t, a_t) can be continuous or discrete.

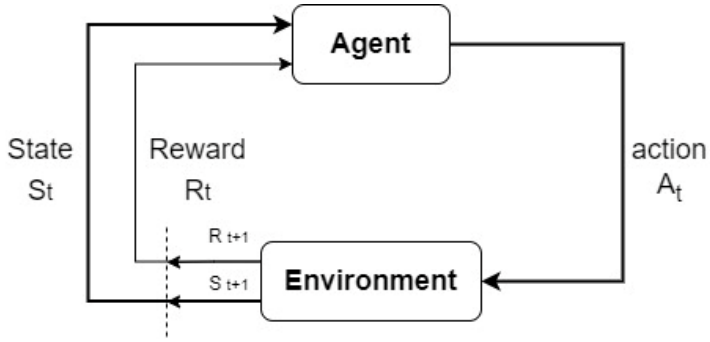

Figure 1. Markov decision process (adopted from [60]).

The goal of an agent is to maximize its cumulative discounted return of rewards G_t as given in (1), where γ is the discount factor $\gamma \in (0,1)$ for which the future rewards are discounted. The policy π that the agent follows dictates the actions that the agent takes as a function of the agent's state. A policy π can be either deterministic or stochastic. The expected return, designated as the "on-policy" action-value function for following the policy π, is given as in (2). The optimal action-value function for following the optimal policy π^* is given in (3) by means of the Bellman equation. In RL an agent must follow the optimal policy π^* to maximize its expected discounted reward.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

$$Q^{\pi}(s,a) = \mathbb{E}\left[G_t | S_t = s, \ A_t = a\right] \tag{2}$$

$$Q^*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} Q^*(s',a')\right] \tag{3}$$

3.2. Artificial Neural Networks

Artificial neural networks (ANNs) are inspired by the working mechanism of biological neurons in the human brain to process information and derive meaning from complicated or imprecise data. An ANN is comprised of an input layer, one or more hidden layers and an output layer. Each neuron in these layers is interconnected to another with an associated weight and a threshold. The capability of ANNs to act as universal function approximators is widely utilized in many reinforcement learning algorithms in which the relationship between dependent and independent variables is not clearly understood [61,62]. There are various types of ANNs for different purposes. Some of the most commonly used ANNs in power system applications are deep neural networks (DNN), convolutional neural networks (CNN), recurrent neural networks (RNN) and graph neural networks (GNN).

3.2.1. Deep Neural Networks

A neural network with more than one hidden layer between the input layer and the output layer is called a deep neural network (DNN). The number of layers and total neurons depend on the complexity of the function. DNNs contain fully connected layers, otherwise known as dense layers, in which every neuron in a particular layer is connected to all the neurons in the next layer. The inputs that are fed to the input layer are multiplied with weights and fed to the activation function. The primary function of the activation function is to add nonlinearity to the neural network by transforming the weighted input sum received to a particular neuron to an output value and passing it to the next neuron. The weights of the DNN are updated during training using backpropagation which computes the gradient of the loss function with respect to weights [63,64]. Successful training of DNNs is dependent on several factors and [65] provides the theory for training ANNs and an overview of various optimization algorithms.

3.2.2. Convolutional Neural Networks

Convolutional neural network (CNN) is a class of ANNs that specialize in pattern detection and is commonly applied to analyse visual imagery. A CNN typically consists of three main types of layers: convolutional layer, pooling layer and fully connected layer. The convolutional layer is a type of hidden layer that carries out convolution operations. Each neuron in a convolutional layer is defined by a low-dimensional matrix that convolves with the input matrix of a higher dimension which leads to an output matrix that is passed to the next layer. A convolutional layer is typically followed by additional convolutional layers or pooling layers. A pooling layer reduces the number of parameters in the input and conducts dimensionality reduction. Although information is lost in the pooling layer, it improves efficiency by reducing complexity and limiting the risk of overfitting. The fully connected layers connect to the output layer and perform the task of classification based on the features extracted from the previous layers. CNNs are widely utilized in power system applications such as network transient stability assessments [66], fault identification [67] and assessment of power quality disturbances [68].

3.2.3. Recurrent Neural Networks

Recurrent neural networks (RNNs) are one of the popular forms of ANNs that specialize in processing long, sequential data or time series data. RNNs are distinguished by their memory function in which the information from prior inputs is utilized to generate the next output of the sequence. Contrary to DNNs that assume the inputs and outputs are independent of each other, the output of the RNNs depends on the prior elements within the sequence. There are different variations of RNNs, such as bidirectional recurrent neural networks (BRNN) [69], long short-term memory (LSTM) [70] and gated recurrent units (GRUs) [71]. RNNs are widely used in power system applications such as photovoltaic power forecasting [72,73] and HC analysis [7,9] that generally require the processing of time series data.

3.2.4. Graph Neural Networks

Graph neural networks (GNNs) are special types of ANNs that are designed to learn from a graph data structure. A typical GNN constitutes of data points called nodes which are linked by lines known as edges. The core function of GNNs is to learn nodal embeddings using the message-passing mechanism, where the features of the nodes are based on the learnable parameters that transform the messages and features of the neighbouring nodes. The learned nodal embeddings are aggregated and passed through a readout function, which is typically a dense layer that outputs the final prediction. Different GNN variants such as graph convolutional network (GCN), graph attention network (GAT) and graph recurrent network (GRN) have demonstrated excellent performances on various deep learning applications [74]. In power system applications, GNNs are typically leveraged to extract topological information in power distribution networks for reinforcement learning algorithms. Some of the literature that utilizes GNNs to solve various control problems can be found in the fields of Volt-VAr control [75], stability control [76] and active-reactive power coordination [77].

4. A Classification of Reinforcement Learning Algorithms

A taxonomy and categories of the most popular RL algorithms are provided in [78], and RL algorithms are divided into two main categories: model-based and model-free. The model-based RL methods require or learn the model of the environment and the model-free methods do not model the environment but look for the optimal policy directly. RL algorithms used for coordinated voltage control in most studies are model-free. The reason is the difficulty of guaranteeing a good model of the environment, which is the electricity distribution network, due to the uncertainties in its network topology, DERs and loads. Model-free RL algorithms can be further categorized into value-based and policy-based. Value-based RL algorithms such as Q-learning derive the optimal policy by determining

the optimal action-value function $Q^*(s,a)$, and policy-based RL algorithms optimize the policy directly. A classification of the most commonly used types of RL algorithms for coordinated voltage control is given in Figure 2.

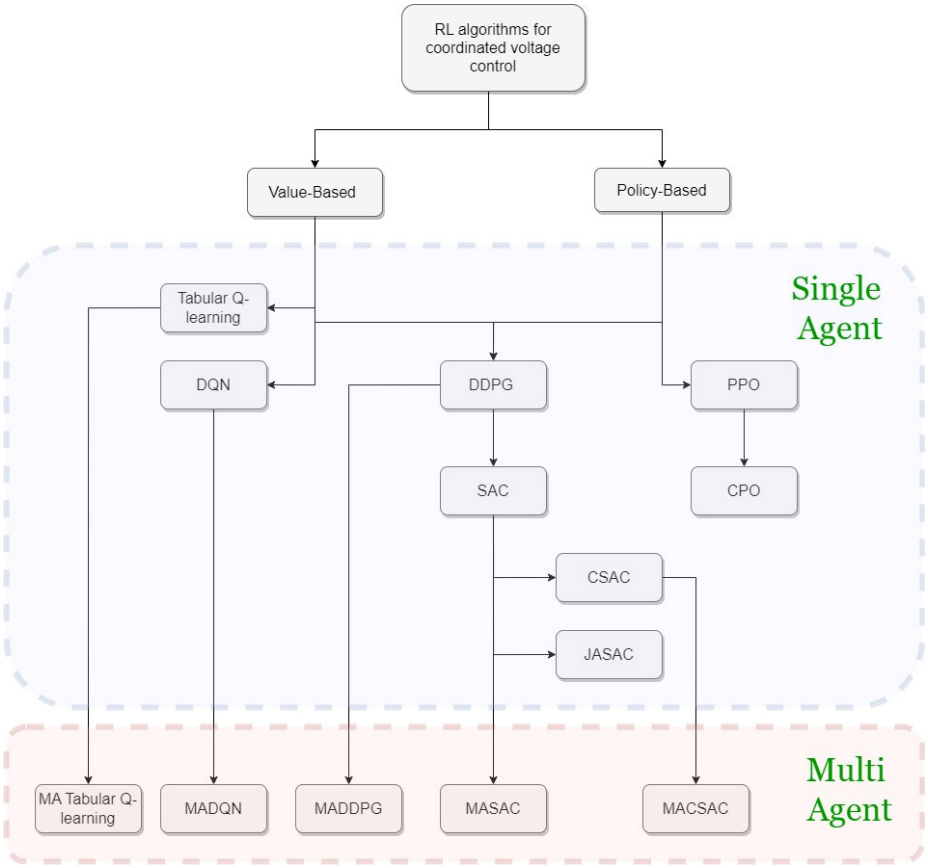

Figure 2. RL algorithms for coordinated voltage control.

4.1. Value-Based Methods

Value-based methods indirectly find the optimal policy by determining the optimal action-value function $Q^*(s,a)$. Common RL value-based methods used for coordinated voltage control include Q-learning and deep Q network (DQN) and its variants. Value-based methods have high sample efficiency, as past experience samples are used for policy updates. As the variance of value function estimation is small, convergence is guaranteed in value-based methods, as it does not easily fall into local optimum. However, value-based methods are only suitable for discrete action spaces and suffer from overestimation bias.

4.1.1. Q-Learning

Q-learning algorithms are model-free algorithms that estimate the expected return of an action and form a given state s. The estimated return for a particular state–action pair is known as the Q-value for that particular state–action pair. In Q-learning, it is desired that the Q-value for a given state–action pair $Q(s,a)$ eventually converges to its optimal Q-value Q^*. An action a in state s is considered to yield better results if its Q-value is high. Q-values are learned and updated iteratively according to (4), where α is the learning rate $\alpha \in (0,1)$. Q-values for each state–action pair are generally stored in a table known as the Q-table, and the dimensions of the Q-table are dependent on the number of actions and states. In [79], a tabular Q-learning algorithm is proposed for the reactive power control in distribution systems. The action spaces of the controllable devices (OLTCs and reactive

power compensation devices) in this study are discretized, since Q-learning can only be applied to discrete action spaces.

$$Q^{new}(s,a) = Q(s,a) + \alpha\left(R_{t+1} + \gamma \max_{a'} Q(s',a') - Q(s,a)\right) \quad (4)$$

4.1.2. Deep Q-Learning

In deep Q-learning, the agent is replaced with a deep neural network that estimates the Q-values for each state–action pair and approximates the optimal Q-function $Q^*(s,a)$. This deep neural network that approximates the Q-function is known as the deep Q network (DQN). Similar to Q-learning, DQN algorithms can only be applied to discrete action spaces. To train a DQN, "experience replay" is used, in which the training data are sampled randomly as a mini-batch from a set of past experiences stored in the "replay buffer". This breaks the temporal correlations among the samples used in the training process and improves the stability of Q-learning. Sampling from previous experiences also increases the data efficiency of the training process. In deep Q-learning, to further mitigate the instability of training, a second neural network called the "target network" is used to estimate the $Q(s',a')$ term in the Bellman equation. The parameters of the "target network" are periodically updated with the parameters of the main network which has been trained.

The *max* operator in DQN uses the same values to select and evaluate an action. This makes DQN more prone to select overestimated values, resulting in overoptimistic value estimates. The double-DQN variation of deep Q-learning decouples the action selection and evaluation to reduce overestimations. This is achieved by learning two value functions, one to determine the greedy policy and another to determine its value. The online network is used to evaluate the greedy policy and the "target network" of DQN is utilized to estimate its value, without introducing any additional networks [80]. Some of the other variations of DQN are Dueling DQN [81], NoisyNet DQN [82] and Distributed DQN [83]. In [84], a two-timescale voltage control method is proposed to control capacitor banks and reactive power output of smart inverters using a DQN-based algorithm. Furthermore, it is demonstrated that tabular Q-learning-based algorithms become less feasible as the dimensionality of the action space increases, and DQN-based algorithms address this issue by providing compact low-dimensional representations of high-dimensional inputs.

4.2. Policy-Based Methods

In contrast to the value-based methods, policy-based methods optimize the policy directly. Policy-based methods can be applied to continuous or higher-dimension action spaces with the advantage of simpler policy parameterization. Policy-based RL algorithms such as Trust Region Policy Optimization (TRPO) [85] and Proximal Policy Optimization (PPO) [86] update their policies in an "on-policy" manner, in which new samples are generated to update the policy by following the current policy. Such methods are considered to be more stable but suffer from low sample efficiency, as they are "on-policy" algorithms. In contrast, value-based based RL algorithms such as DQN are "off-policy" algorithms, in which the samples for the policy update are generated by following a different policy called "behavioural policy". Vanilla policy gradient methods aim to directly learn the optimal policy by estimating the gradient of the policy as given in (5), where π_θ represents a policy with parameter θ and $r(s_t,a_t)$ is the immediate reward at state s_t for taking an action a_t. The policy parameters θ are updated via gradient accent in the direction that provides the maximum expected rewards [87].

$$\nabla J = \sum_{t=0}^{H} \nabla_\theta \log\ \pi_\theta(a_t|s_t)\ r(s_t,a_t) \quad (5)$$

The study presented in [88] formulates the optimal power flow of distribution networks as an MDP and employs the PPO algorithm to solve the MDP sequentially. The main objective of the proposed method is to minimize the cost of power loss while considering

voltage constraints of the network. Such algorithms may be effective for voltage regulation to some extent but will not provide the best solution to regulate the voltage, as it is not their main objective. In [75], the PPO algorithm is combined with a graph convolutional neural network (GCN) to regulate the voltage in power distribution systems. Capacitors, voltage regulators and batteries are the controlled network elements in this study, and the action space of each controlled element is discretized. The graph representation for the GCN for this study is induced by the physical power system topology. A comparative study between dense neural networks (DNN) and GCN used for the PPO algorithm is presented in this study, and the results indicate that both policies converge to the same reward when used in the PPO algorithm with GCN converging at a slower rate. However, the advantage of using GCN is that it is more robust to the errors associated with data misalignments and communication failure.

4.3. Actor-Critic Algorithms

The most commonly used algorithms for coordinated voltage control are actor-critic algorithms such as deep deterministic policy gradient and soft actor-critic, which combine the merits of value-based and policy-based methods. These algorithms possess the advantages such as high sample efficiency from value-based methods and the applicability to both continuous and discrete action spaces from policy-based methods. However, these actor-critic algorithms also inherit several disadvantages, such as overestimation bias and insufficient exploration from the merge of value-based and policy-based methods. In actor-critic algorithms, two functions actor and critic are parameterized by neural networks. The critic estimates the action value (Q) or the state value (V) and the actor updates the policy distribution suggested by the critic with policy gradients.

4.3.1. Deep Deterministic Policy Gradient

One of the drawbacks of Q-learning is that it cannot be straightforwardly applied to continuous action spaces because it is computationally expensive to exhaustively evaluate a continuous action space and use a normal optimization algorithm to calculate $\max_{a'} Q^*(s,a)$. Hence, an actor-critic approach based on the deterministic policy gradient algorithm (DDPG) is proposed in [89] for reinforcement learning with continuous action spaces. Deterministic policy gradient exploits the differentiability of $Q^*(s,a)$ with respect to the action argument in continuous action spaces to establish a gradient-based policy rule for a policy $\mu(s)$.

DDPG is an actor-critic method that uses deep neural networks to parameterize policies (θ) and concurrently learn a Q-function and a policy. The parameterized actor function $\mu(s|\theta^\mu)$ specifies the current policy in DDPG by deterministically mapping states to a specific action. The critic function is learned similar to Q-learning by using the Bellman equation and the off-policy data which are generated from a different stochastic behaviour policy β. The critic parameters are updated by minimizing this mean-squared Bellman error loss function according to (6) and (7). The actor is updated according to (8), which is the derivative of the start distribution J with respect to the actor parameter θ^μ and it is simplified further by applying the chain rule [90]. Similar to DQN, a replay buffer and target networks are used in DDPG to improve the stability of training.

$$L\left(\theta^Q\right) = \mathbb{E}_{s_t\sim\rho^\beta,\ a_{t\sim\beta},\ r_t\sim E}\left[\left(Q\left(s_t,a_t\middle|\theta^Q\right)-y_t\right)^2\right] \tag{6}$$

where

$$y_t = r(s_t,a_t)+\gamma Q\left(s_{t+1},\mu(s_{t+1})\middle|\theta^Q\right) \tag{7}$$

$$\begin{aligned}\frac{\partial J(\theta^\mu)}{\partial\theta^\mu} &= \mathbb{E}_{s_t\sim\rho^\beta}\left[\left|\nabla_{\theta^\mu}Q(s,a|\theta^\mu)\right|_{s=s_t,\ a=\mu(s|\theta^\mu)}\right]\\ &= \mathbb{E}_{s_t\sim\rho^\beta}\left[\left|\nabla_a Q\left(s,a\middle|\theta^Q\right)\right|_{s=s_t,\ a=\mu(s_t)}\left|\nabla_{\theta_\mu}\mu(s,a|\theta^\mu)\right|_{s=s_t}\right]\end{aligned} \tag{8}$$

DDPG algorithms are widely used by many authors to control various network elements such as PV systems, static VAr compensators (SVCs), battery energy systems, flexible loads and smart transformers to regulate the voltage in the distribution network [71,91,92]. The inverter-based DERs in the electricity distribution network are fast timescale devices that can be controlled within seconds. DDPG is well suited for such control problems with continuous devices due to its applicability to continuous state-action spaces. A comparison between the double-DQN algorithm and DDPG algorithm applied to the same voltage control problem is provided in [91]. The results indicate that DDPG achieves a better performance of the two, as double-DQN is incapable of fully utilizing the voltage control capabilities of the controlled elements due to the aggregation and discretization of actions. Several key factors that impact the performance of DDPG algorithms used for coordinated voltage control are listed as follows:

- The replay buffer size—The performance of DDPG increases with the size of the replay buffer, but converges after a certain value. Identification of the correct replay buffer size is key in maximizing the performance of DDPG while utilizing less memory.
- Exploration strategy—DDPG algorithms are deterministic and do not explore the environment as do other stochastic algorithms, and they often converge to a local optimum. One of the strategies to achieve exploration in DDPG algorithms is to add Gaussian noise to the actions during the training process.
- Accuracy of the training model—DDPG algorithms are often trained offline before their implementation in the online distribution network to minimize the undesirable costs associated with the start-up stage. Therefore, the distribution network model used for training must be accurate to achieve a good level of performance and low costs during the online execution.

It is not always possible to have an accurate model of the distribution network, especially the LV distribution network, due to inaccuracies and unavailability of data on network elements and network topology. One of the solutions proposed in recent literature to this dilemma is to use a data-driven DNN-based surrogate model that estimates the bus voltages and power losses in the distribution network [91,92]. Training samples for DDPG are generated by the DNN-based surrogate model and control policies are learned offline, without interacting with the actual distribution network. However, it is difficult for a DNN-based network model to estimate line parameters and the topology of the actual distribution network. Changes to the distribution network topology such as the increase or decrease in network nodes or branches may produce undesirable results in the execution phase of DDPG control policies that are trained with a DNN-based surrogate model.

In [77], a graph attention network (GAT) is leveraged to extract and aggregate topology branch correlations and node power information. The extracted topological information is embedded with node features and fed into the underlying network architecture of the DDPG algorithm as intermediate latent environment states. The attention mechanism of GAT aggregates neighbour nodes and allocates weights adaptively according to the correlation of node information in feature extraction. The proposed method regulates the voltage of the distribution network using a real-time coordinated scheduling scheme of active and reactive power of controllable devices such as PV systems, flexible loads, energy storage systems and SVCs. The use of GAT embedded into DDPG can achieve highly desirable performance levels even under unknown electrical fault scenarios and topology variations.

Since DDPG-based algorithms can only be applied to continuous action spaces, they cannot be used to control network elements with discrete action spaces such as OLTCs, voltage regulators and capacitor banks. This can be seen as a disadvantage of using DDPG-based algorithms for optimal control operations in the distribution network.

4.3.2. Soft Actor Critic

DDPG algorithms train a deterministic policy in an off-policy manner and are highly sample efficient, as the policy update uses past experience samples in the replay buffer.

However, DDPG is considered to be brittle due to its sensitivity to hyperparameters and often requires more tuning to converge. Soft actor-critic algorithms (SACs) are considered to be more robust to hyperparameters and more stable than DDPG algorithms, as they optimize a stochastic policy in an off-policy way. SACs introduce entropy regularization, in which the policy is trained to maximize both its expected return and entropy. The randomness of the policy is termed as entropy, and it influences the exploration–exploitation trade-off.

$$\mathrm{J}(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))] \tag{9}$$

The objective function of an SAC is given in (9), where $\mathcal{H}(\pi(\cdot|s_t))$ is the entropy term weighted by the temperature parameter α. Maximizing entropy encourages the policy network to explore, which accelerates learning and prevents the policy from converging into a local optimum. SAC utilizes three parameterized networks: state-value function $V_\psi(S_t)$, soft Q-function $Q_\theta(s_t,a_t)$ and a policy function $\pi_\phi(a_t|s_t)$. The parameters for the state-value function, soft Q-function and policy function are ψ, θ and ϕ, respectively. SAC is implemented on continuous action spaces similar to DDPG. However, it is also possible to implement SAC on discrete action spaces by slightly changing the policy update rule. The parameters for the state-value function, soft Q-function and policy function are updated according to (10), (11) and (12), respectively [93]. A target state-value function $V_{\overline{\psi}}(S_{t+1})$ is used for the soft Q-function update and to ensure training stability. For the policy update in (12), the reparameterization trick is employed, where $a_t = f_\phi(\epsilon_t;s_t)$ and ϵ_t is an input noise vector.

$$\hat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(S_t)\big(V_\psi(S_t) - Q_\theta(s_t,a_t) + \log\ \pi_\phi(a_t|s_t)\big) \tag{10}$$

$$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(s_t,a_t)\Big(Q_\theta(s_t,a_t) - r(s_t,a_t) - \gamma V_{\overline{\psi}}(S_{t+1})\Big) \tag{11}$$

$$\hat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi \log\ \pi_\phi(a_t|s_t) + \big(\nabla_{a_t}\log\ \pi_\phi(a_t|s_t) - \nabla_{a_t}Q(s_t,a_t)\big)\nabla_\phi f_\phi(\epsilon_t;s_t) \tag{12}$$

In addition to replay buffers, target networks and entropy regularization, one of the crucial tricks employed by SAC algorithms is the clipped double-Q learning, which learns two Q-functions and uses the smaller Q-value of the two in the Bellman error loss function. This facilitates fending off the overestimation bias in the Q-function. In [94], an SAC algorithm is used to regulate the voltage by regulating the inverter-based devices and SVCs in the distribution network. The proposed SAC-based control framework is safe, stable and efficient, as the control policies are trained offline utilizing historical operational data before online execution in the actual distribution network. One of the advantages of SAC over other algorithms such as DDPG is its ability to learn control policies for both continuous and discrete action spaces. Ref. [95] is one of such applications of SAC in voltage regulation, where control policies are learned for both slow timescale devices (OLTCs and capacitor banks) with discrete action spaces and fast timescale devices (PV inverters and SVCs) with continuous action spaces. In this study, two separate SAC agents were utilized, each for slow timescale devices and fast timescale devices, and through coordinated control between agents, stable and satisfactory optimized voltage control is achieved.

Certain exploratory control actions produced by various RL algorithms including SAC may lead to significant voltage violations in the distribution network that undermine the reliability of the network. In most studies, network constraints are implemented by simply augmenting the reward with a penalty term. This often makes the learned policy of the RL algorithm infeasible or too conservative. In contrast, RL algorithms formulated on the Constrained Markov Decision Process (CMDP) for coordinated voltage control often achieve near-perfect constraint satisfaction.

CMDP can be defined as a tuple $(S,\ A,\ p,\ r,\ c,\ d,\ \gamma)$, similar to MDP, where an agent interacts with the environment at discrete time step t at state $s\ \epsilon\ S$ and then takes an action $a\ \epsilon\ A$ to receive a reward $r(s,a)$ and a cost $c(s,a)$; p is the transition probability function, γ is the discount factor $\gamma\ \epsilon\ (0,1)$ and d is the safety threshold. The goal of the agent in CMDP

is given in (13), where an agent learns a policy π that maximizes the expected return such that the safety constraint violations remain below the threshold d for each time step t [96].

$$\max_{\pi} \ \mathbb{E}_{(s_t,a_t)\sim p_\pi}\left[\sum_t \gamma^t \ r(s_t,a_t)\right] s.t. \ \ \mathbb{E}_{(s_t,a_t)\sim p_\pi}\left[\sum_t \gamma^t \ c(s_t,a_t)\right] \ \leq \ d \tag{13}$$

The SAC algorithm, which has been demonstrated to successfully implement coordinated voltage control, cannot be used to solve CMDPs, as it is designed to solve MDPs. The constrained soft actor-critic (CSAC) algorithm extends SAC to satisfy the operational constraints in CMDP and solve the constrained optimization problem by using the Lagrange-multiplier method [97]. The CSAC algorithm employs two critics: reward critic and safety critic. The reward critic is trained to express the estimation of long-term rewards with entropy and the safety critic is trained to express the estimation of long-term costs to encourage safety. The trade-off between reward and safety is managed by using adaptive safety weights.

Several variations of the CSAC algorithm are used in the recent literature to solve the CMDP and learn a parameterized control policy to regulate the voltage in the electricity distribution network [98,99]. The use of CSAC in voltage control enables safe exploration for controllable devices and guarantees operational constraint satisfaction in the form of expectation. However, direct online implementation of CSAC in the distribution network may result in some voltage violations during the start-up stage that could lead to poor network reliability and an increase in economic costs. This is due to the learned control policy of the CSAC algorithm being weak during the initial start-up stage. In [98], the performance of CSAC on the same voltage control problem is compared with other algorithms such as SAC, DQN and constrained policy optimization (CPO), and the results indicate that CSAC achieves better performance in relation to sample efficiency, scalability and constraint satisfaction.

Joint adversarial soft actor-critic (JASAC) is another variation of SAC proposed in [100] to regulate the voltage of distribution networks by controlling the reactive power of the inverter-based energy resources and SVCs. The JASAC algorithm formulates the Volt-VAr control problem as an adversarial Markov decision process (AMDP), which is an extension of the MDP. A protagonist and an adversary are involved in the AMDP and the goal of the adversarial agent is to hinder the protagonist by adjusting for the network modelling errors. The proposed method is implemented in two stages: offline stage and online stage. In the offline stage, a model of the distribution network is used in the training of an offline agent using JASAC, and in the online stage, the offline agent is transferred to the online agent to perform continuous learning and control using the SAC algorithm on the actual distribution network. Performance of the JASAC algorithm is benchmarked against SAC for the same voltage control problem and the results indicate that the JASAC performs better than SAC in the online stage due to its robustness to changes in network parameters.

5. Multi-Agent Reinforcement Learning

In multi-agent reinforcement learning (MARL) a number of individual agents communicate with each other and interact with the environment to solve complex tasks. The sequential decision-making process of MARL is formalized using Markov games, which is an extension of the MDP. A Markov game at time step t for K number of agents can be defined by the tuple $\left(S,\ \{O^i\}_{i=1}^K \{A^i\}_{i=1}^K,\ P,\ \{r^i\}_{i=1}^K\right)$ that consists of a global state S, K number of local observations O^i for each agent i, K number of local action spaces A^i for each agent i, a global transition probability P and K number of local reward functions r^i for each agent i. For each discrete time step t, each agent selects a local action A_t^i based on the local observation O_t^i. Then, each agent receives a reward $R_{t+1}^i = r^i\left(S_t, A_t^1, A_t^2, \ldots, A_t^K\right)$ as a function of the state S_t and the joint action $A_t = \left[A_t^1, A_t^2, \ldots, A_t^K\right]$. Consequently, the environment's global state transitions to S_{t+1} based on the state transition probability $P\left(S_{t+1}\middle|S_t, A_t^1, A_t^2, \ldots, A_t^K\right)$ and each agent receives next local observation O_{t+1}^i. All agents

communicate with each other at every time step and share local information to identify the local control policy such that the joint policy π of all the agents maximizes the expected discounted average return $J(\pi)$ as given in (14) [101].

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \frac{1}{K} \sum_{i=1}^{K} R_{t+1}^{i}\right] \tag{14}$$

A direct extension of single-agent RL into a multi-agent environment as an independent Q-learning algorithm will be computationally expensive and vulnerable to overfitting, as each agent aims to learn an independent policy by assuming other agents as a component of the environment. An alternative and the standard paradigm of multi-agent RL is the centralized learning and the decentralized execution approach. A multi-agent tabular Q-learning algorithm is proposed in [102] to regulate the voltage of the distribution network through centralized learning of control policies and decentralized execution of control actions. Each controllable network element is assigned to an agent that maintains and updates a Q-table during the learning process. The optimal control action for each controllable network element is determined based on the Q-values of the respective agent's Q-table. However, for large control problems, multi-agent DQN (MADQN) methods are always preferred over traditional multi-agent tabular Q-learning methods, as they become less feasible with the increase in dimensions of the Q-table.

In [57], a decentralized voltage control method is proposed using a MADQN algorithm to control network elements such as smart inverters, switchable capacitors and voltage regulators. MADQN algorithms are more scalable than single-agent DQN, as the action space of single-agent DQN increases exponentially with the increase in controllable elements. The training of the proposed RL algorithm is performed offline in a centralized manner and online execution is performed in a decentralized way. During the training process of the proposed MADQN algorithm, each agent takes an action based on their local observations and the action is evaluated by considering the overall Q-value of combined actions of all agents. One of the advantages of the decentralized execution of MADQN policies is that it involves no communication constraints. However, the convergence theory of single-agent Q-learning that is extended to MARL is not guaranteed in the presence of a non-stationary environment. The environment becomes non-stationary if the learning among agents constantly reshapes the environment and affects the optimal policy of agents. The exploration–exploitation trade-off could be more influential in a multi-agent setting and the interactions among agents must be performed while ensuring the stability of the agents [103].

Multi-agent deep deterministic policy gradient (MADDPG) algorithms adopt a framework of centralized training with decentralized execution for continuous action spaces. The MADDPG algorithm is a simple extension of the DDPG where the agents learn a centralized critic, which is augmented with the policies of all the agents. However, the actor only has access to local information. MADDPG is applicable for cooperative settings due to the capability of agents to learn the policies of other agents online and incorporate them in their own policy-learning procedure. According to [104], actor and critic parameters of MADDPG are updated according to (15) and (16, respectively, where X represents the state information and $\mathcal{D}$ is the experience replay buffer that contains the experience tuples $(X, X', a_1, \ldots, a_N, r_1, \ldots, r_N)$ of all agents.

$$\nabla_{\theta_i} J(\mu_i) = \mathbb{E}_{X,a\sim\mathcal{D}}\left[\left.\nabla_{\theta_i}\mu_i(a_i|o_i)\nabla_{a_i}Q_i^{\mu}(X, a_1, \ldots, a_N)\right|_{a_i=\mu_i(o_i)}\right] \tag{15}$$

$$L(\theta_i) = \mathbb{E}_{X,a,r,X'}\left[\left(Q_i^{\mu}(X, a_1, \ldots, a_N) - y\right)^2\right] \tag{16}$$

$$y = r_i + \gamma Q_i^{\mu}(X', a_1', \ldots, a_N')\Big|_{a_j'=\mu_j'(o_j)'} \tag{17}$$

In [105], a two-stage control scheme is proposed to control the network elements such as OLTCs, capacitor banks and PV systems to regulate the voltage in the distribution network. In the first stage, a day-ahead dispatch of OLTC and capacitor banks is obtained by solving the optimal power flow problem using mixed-integer second-order cone programming. In the second stage, the MADDPG algorithm is used to control the reactive power outputs of PV systems using the dispatched results for the OLTCs and capacitor banks in the first stage. The objective of the proposed MADDPG algorithm is to minimize the network power loss while satisfying the voltage violation constraints and PV systems' reactive power capability constraints. Due to the decentralized online execution of the proposed MADDPG algorithm, deployment of costly communication infrastructure is not required. However, the performance of the proposed MADDPG relies on an accurate network model for offline training. Therefore, its performance may degrade in the instances of network expansions and network reconfigurations that change the network model.

In [106], an autonomous voltage control framework is proposed using the MADDPG algorithm. The objective of the proposed algorithm is to maintain the voltage bus magnitudes within the desirable levels. The critic of the proposed MADDPG algorithm treats actions of all agents equally. This means that the spatial properties between agents are ignored and may lead to performance degradation when applied to a large system with a high number of control elements (agents), as the input of the critic increases with the number of agents. An attention-critic-based MADDPG method is proposed in [107] to enhance the scalability of the MADDPG algorithm used for the coordinated voltage control of PV inverters. The attention critic allows for the intelligent learning of specific information that is most relevant to the rewards when the number of control objects are high, especially in large distribution networks.

A variation of the multi-agent soft actor critic (MASAC) algorithm is proposed in [101] to control voltage regulators, capacitor banks and OLTCs to regulate the voltage in a decentralized manner. Communication efficiency is achieved, as each agent maintains a local replay buffer and information is transmitted only within the neighbouring agents. The proposed algorithm is identified to have the same performance level as that of a single-agent SAC algorithm, but the key difference between the proposed algorithm from centralized algorithms such as SAC is that the system continues to function even when there is a communication link breakdown, as it is implemented in a decentralized manner. However, if the MASAC algorithms are directly implemented on the distribution network, the exploratory actions in the early stages of learning may cause constraint violations and degradation of switching elements. These exploratory actions of MASAC will diminish over time but a few will persist due to entropy regularization.

To achieve excellent constraint satisfaction and reduce communication delay during training, a variation of the multi-agent constrained soft actor critic algorithm (MACSAC) is proposed in [108], which controls PV inverters and SVCs in the distribution system to regulate the voltage. The control policies of the implemented MACASC algorithm are learned online in a centralized manner and executed in a decentralized manner by the local controllers. Each agent communicates with a control centre asynchronously, where a centralized critic is learned considering observations of all the agents. The latest control policies are then sent to the local controllers by the control centre. The implemented asynchronous learning, sampling and control process does not create any delay to the control actions due to the training process and it is highly efficient in terms of communication. The performance of MACSAC is compared against CSAC and MADDPG and the results demonstrate that MACSAC has similar performance to CSAC but higher performance when compared with MADDPG. However, MACSAC is efficient in communication when compared with CSAC due to its decentralized execution of control actions.

6. Environment Model and MDP Parameters

This section discusses several key components of the recently proposed RL algorithms for coordinated voltage control including the environment model, state space, action space, reward function and constraints.

6.1. Environment

The electricity distribution network is the environment of the coordinated voltage control problem solved through reinforcement learning methods. However, most of the references rely on accurate electricity distribution network models to test and train the proposed algorithms. This is due to the risks associated with implementing untrained RL algorithms in real distribution networks that could potentially lead to unbearable costs and jeopardize network reliability. Offline training of RL agents requires an accurate model of the electricity distribution network to achieve higher performance levels. Developing a good simulation model often requires an extensive amount of historical data and accurate information on network topology and network elements. For the instances in which such data are not available, some studies model the distribution network in the form of DNN [91,92] and GNN [75,77] instead of the traditional power flow models.

Distribution networks such as the IEEE 123-bus system, IEEE 33-bus system and IEEE 37-bus system are some of the most commonly used distribution network test environments in the recent literature. A distribution network with a certain degree of unbalance is more suitable as a test environment for the proposed RL algorithms, since the voltage control actions are sensitive to the unbalance of the distribution network. It should be noted that the majority of the proposed RL algorithms are tested on MV distribution networks. However, the voltage control problem is more prominent in low-voltage (LV) networks with low reactance/resistance ratio of the conductors, and testing the proposed RL algorithms on LV networks will give a good indication of their performance.

6.2. State Space

A state is a representation of the environment that influences the decision-making process of an agent, reward calculation and transitions of the environment. The state space of an MDP is the set of all possible states in an environment. The voltage control problem in the distribution network depends on a range of network variables. A state representation that consists of much redundant information may slow down the learning process. Therefore, incorporating the most relevant variables to the voltage control problem in the state information is key to achieving a higher level of performance for any RL algorithm. Some of the distribution network variables featured in the state space of proposed RL algorithms in recent literature are as follows:

- Active and reactive power of customer loads [109];
- Active and reactive power injections of PV systems [110];
- Three-phase voltages of buses [111];
- Tap positions of OLTCs, capacitor banks and voltage regulators [112,113].

The current state information of the environment makes up the states in most of the literature. However, it is possible to incorporate past experiences in the state information as in [94], which includes the previous action in the states. The distribution network topology cannot be easily represented by the state values in an MDP, and a solution to this is to use GNN-based policies that capture the topological information of the network as inputs to the RL algorithm [75,77].

6.3. Action Space

The action space is the set of all possible actions that an agent can take in a specific environment. In the coordinated voltage control problem, the actions of an RL agent are typically the control actions of the voltage-regulating devices in the distribution network. In many cases, these actions are state dependent and can be discrete or continuous depending

on the controlled device. Some of the controlled devices and their respective actions of RL algorithms proposed in the recent literature are as follows:

- Reactive power of PV inverters [105,107];
- Active power curtailed by PV systems [91];
- Discretized actions for the tap positions of voltage regulators, capacitor banks and OLTCs [101];
- Reactive power output of SVCs [108];
- Reactive power adjustments of the smart transformers [92].

One of the instances in which the action space of a RL algorithm does not belong to a controllable device can be found in [100], where modelling errors related to line reactance and resistance are considered as actions in the proposed JASAC algorithm. A recurring challenge in RL algorithms with discrete actions is that the size of the action space increases exponentially with each additional feature in the state. This is commonly known as the "curse of dimensionality", and one of the solutions to this dilemma is presented in [98], where the policy network is designed with a device-decoupled structure to ensure the network structure increases linearly with the number of controllable devices.

6.4. Reward Function and Constraints

The reward function determines the direct reward received by an agent. The objectives in solving the problem are reflected in the formulation of the reward function. Incorrect formulation of the reward function will result in undesired behaviour of RL algorithms. The reward function for the voltage control problem is formulated differently in various literature. In [91,107], the reward function is formulated to reduce the voltage deviations throughout the network from its nominal value. This makes voltage control the main objective of the RL agent. However, most of the proposed RL algorithms consider voltage constraints as a penalty factor embedded in the reward function along with several other objectives as listed below:

- Reduce the active power loss in the network [114];
- Reduce network operational costs [113];
- Reduce device switching costs [101].

Other methods for implementing voltage constraints are by means of a safety layer that operates on top of the RL algorithm [92] and explicitly modelling voltage constraints as a cost function [98,108].

7. Risk Factors and Challenges

This section presents the centralized and decentralized facets of RL-based coordinated voltage control methods and discusses some of the critical challenges of implementing such algorithms in the electricity distribution network. A comparison of several RL-based algorithms considering key challenges that are associated with the voltage control problem, such as scalability, sample efficiency, robustness to network changes, constraint satisfaction and safety, are provided in Table 2. Furthermore, a summary of references for RL-based coordinated voltage control and their key features is provided in Table 3.

Table 2. Comparison of RL algorithms used for coordinated voltage control.

Algorithm	Ref.	Sample Efficiency	Scalability	Constraint Satisfaction and Safety	Robustness to Network Changes	Centralized/ Decentralized
Tabular Q-learning	[79]	High	Low	Low	Low	Centralized
DQN	[84]	High	Low	Low	Low	Centralized
PPO	[75,88]	Low	High	Medium	Low	Centralized
DDPG	[77,91,92]	High	High	Medium	Low	Centralized

Table 2. *Cont.*

Algorithm	Ref.	Sample Efficiency	Scalability	Constraint Satisfaction and Safety	Robustness to Network Changes	Centralized/ Decentralized
SAC	[94,95]	High	High	Medium	Low	Centralized
CSAC	[98,99]	High	High	High	Low	Centralized
JASAC	[100]	High	High	High	High	Centralized
MA Tabular Q-learning	[102]	High	High	Low	Medium	Decentralized
MADQN	[57]	High	High	Low	Medium	Decentralized
MADDPG	[105,106,111]	High	High	Medium	High	Decentralized
MASAC	[101,110,112]	High	High	Medium	High	Decentralized
MACSAC	[108]	High	High	High	High	Decentralized

Table 3. A summary of references for RL algorithms used for coordinated voltage control.

Ref.	RL Algorithm	State Space	Action Space	Algorithm(s) Used for Benchmarking	Implemented Network(s)	Description
[79]	Tabular Q-learning	Constraint violations of busbars	Discretized actions for OLTCs and reactive power compensation devices	Probabilistic constrained load flow and genetic algorithm	IEEE 14-bus system and IEEE 136-bus system	A tabular Q-learning algorithm is proposed to provide offline control settings while satisfying operational limits of the constraint variables
[84]	DQN	Active power of busbars and current capacitor configurations	Discretized actions for capacitor banks (on/off)	Nil	IEEE 123-bus system and a real-world 47 bus network	A two-timescale voltage control algorithm is proposed with slow-timescale learning for optimal capacitor settings using DQN and fast timescale optimization for smart inverter reactive power using optimal power flow models
[75]	GCN-PPO	Minimum-phase voltage at every bus and the control status of voltage regulators, capacitor banks and batteries	Discretized actions for the tap positions of voltage regulators, capacitor banks (on/off) and batteries (charging/discharging)	Dense-PPO	IEEE 13-bus system, IEEE 123-bus system, IEEE 34-bus system and 8500 node system	GCN-based policy network is used in the PPO algorithm to capture the topological information of the distribution network and regulate the voltage
[77]	GAT-DDPG	Connection relations and power information of the distribution network are mapped to graph-structured vertices and edges	Power outputs of PV systems, flexible loads, battery energy systems and SVCs	DDPG GCN-DDPG	IEEE 33-bus system	A GAT-based policy network is combined with DDPG algorithm to optimally schedule controllable devices while being robust to network topology variations
[91]	DDPG	Active and reactive power of customer nodes and active power outputs of PV systems	Active power curtailed in PV systems and reactive power outputs of SVCs and PV systems	Double-DQN, model-predictive control (MPC)	IEEE 123-bus system	A model-free voltage control method is proposed using a DNN-based surrogate model of the distribution network with a DDPG algorithm

Table 3. *Cont.*

Ref.	RL Algorithm	State Space	Action Space	Algorithm(s) Used for Benchmarking	Implemented Network(s)	Description
[92]	DDPG	Voltage magnitudes of bus voltages	Smart transformer reactive power adjustments	MPC	IEEE 33-bus system and IEEE 123-bus system	A DDPG algorithm is utilized with a safety layer technique to achieve the objectives of system power loss reduction and voltage regulation by realizing optimal control policies for smart transformers
[94]	SAC	Voltage, active and reactive power of nodes and the previous action	Reactive power outputs of inverter-based energy resources and SVCs	Without Volt-VAr control, online trained SAC	IEEE 33-bus system and	A PEDNN is utilized to learn the power flow model of the distribution network and then incorporated in the training process of the SAC-based polices that is used to regulate the voltage in the distribution network
[95]	SAC	Active and reactive power injections of buses, voltage magnitudes and tap positions of OLTCs and capacitor banks	Discrete actions for tap positions in the SAC agent controlling slow timescale devices; continuous actions for reactive power injections in the SAC agent controlling fast timescale devices	DQN	IEEE 33-bus system and IEEE 123-bus system	The proposed two-timescale SAC based algorithm learns control policies for both fast timescale and slow timescale network elements to optimize voltage regulation in the distribution network in a coordinated manner
[98]	CSAC	Active and reactive power injections of buses and current tap positions of controlled devices	Discrete actions for tap positions of voltage regulators, OLTCs and capacitor banks.	SAC, DQN and CPO	IEEE 4-bus, 34-bus and 123-bus distribution test feeders	A safe model-free CSAC algorithm is proposed to regulate the voltage in the distribution network and achieve operational constraint satisfaction
[100]	JASAC	Voltage magnitude, active and reactive power of buses in the network	Reactive power outputs of inverter-based energy resources and SVCs	SAC	IEEE 33-bus, 69-bus and 123-bus test feeders	A two-stage deep RL algorithm that is robust to network variations is proposed to regulate the voltage in the distribution network
[102]	MA tabular Q-learning	The local observations for every agent are the active power flows to neighbouring buses	Voltage regulators, capacitor banks and OLTCs are considered as agents and the action space is dependent on the type of controlling device	Discrete particle swarm optimization algorithm and interior point method	Ward-Hale 6-bus system, IEEE 30-bus system and the IEEE 162-bus system	MA tabular Q-learning algorithm is proposed to regulate the voltage of the distribution network and minimize the real power loss while satisfying the operational constraints
[57]	MADQN	Three-phase voltages at all buses in the distribution network	Control actions of smart inverters, autonomous voltage regulators and capacitor banks	Nil.	IEEE 13-bus and 123-bus distribution systems	A MADQN algorithm is proposed to regulate the voltage and reduce power loss in an unbalance distribution network
[105]	MADDPG	Local measurements of PV systems combined with dispatch results of OLTC and capacitor banks	Difference in PV reactive power between two consecutive time steps	One-timescale centralized, double-timescale local and combined centralized and local Volt-VAr	IEEE 33-bus system	A MADDPG algorithm is proposed to regulate the voltage that features offline centralized training and online decentralized execution

Table 3. *Cont.*

Ref.	RL Algorithm	State Space	Action Space	Algorithm(s) Used for Benchmarking	Implemented Network(s)	Description
[106]	MADDPG	States are defined as a vector including system bus voltages, phase angles and power flows	Action space is defined as a vector of generator bus voltage magnitudes	DDPG	Illinois 200-bus system	A data-driven MADDPG algorithm that is robust to a weak communication environment is proposed to solve the autonomous voltage control problem
[107]	MADDPG	A set of local observations consisting of active and reactive power of loads and active power injections of PV systems	The action space is a set of actions by all agents consisting of reactive power outputs of the inverters	QV-droop control and MADDPG without attention critic	IEEE 33-bus system and IEEE 123-bus system	A MADDPG algorithm is developed by incorporating an attention critic model that allows intelligent learning and identifies certain information that is the most relevant to the rewards
[101]	MASAC	State is a set of network active and reactive power injections and the status of the controlled devices in the previous step	Action space is a set of tap positions of OLTCs, capacitor banks and voltage regulators	SAC, MPC and mixed-integer conic programming	IEEE 4-bus, 34-bus and 123-bus distribution test feeders	A randomization-based consensus algorithm is developed utilizing MASAC by establishing communication of each agent with its neighbours to regulate the voltage in the distribution network
[108]	MACSAC	State space is defined as a vector that consists of active and reactive power injections of buses and voltage magnitudes	Action space is defined as vector containing the reactive power outputs of PV inverters and SVCs	CSAC, MASAC and MADDPG	33-bus, 141-bus and IEEE 37-bus distribution test feeders	The distribution network is divided into several areas and each area is given a local controller that act as an agent in the implemented MACSAC algorithm; each agent controls the reactive power of the controllable devices locally in the respective area

7.1. Centralized and Decentralized Control

Most single-agent RL algorithms such as DDPG and SAC regulate the voltage in the distribution network in a centralized manner. Such algorithms require the RL agent to have full access to all the necessary information in the environment. This often requires a centralized controller that constitutes high-performance computers with large-scale data storage backed by a highly reliable communication infrastructure. The single-agent centralized RL algorithm may produce undesirable outcomes in the events of communication breakdown and changes to the network [91,98].

MARL algorithms typically control voltage in the online distribution system in a decentralized manner with learning experiences gained from offline centralized training using a model of the network. In MARL, multiple agents with partial observations of the environment cooperatively learn decentralized policies to achieve a shared objective. This is highly advantageous in solving the voltage control problem in the instances where the decision maker does not have access to all the required information in the environment. Compared to single-agent centralized RL algorithms, the decentralized implementations of MARL algorithms require less communication infrastructure and are robust to communication breakdown [106,107].

7.2. Safety and Scalability

RL-based coordinated voltage control schemes are implemented on top of the vital infrastructure of power distribution networks. Hence, RL algorithms that guarantee the safety of the learning process and learned policies are more desirable. A proposed RL algorithm can be safely implemented in the distribution network if the learned policies satisfy the network operational constraints: robust to communication breakdown and resilient to network variations.

If RL algorithms are directly implemented in the distribution network, it may lead to severe security problems and unbearable costs due to the exploration actions of the RL agents during the start-up phase. Therefore, the RL agents are often preferred to be trained offline before the online implementation in the distribution network. The "transfer gap" is the disparity between the real distribution network and the distribution network model used for the offline training. The trained offline agent may show undesirable performance when transferred to the online agent due to the transfer gap. Therefore, an algorithm that is robust to the network modelling errors is more desirable, since it is very difficult to accurately model all the parameters of the distribution network. Some of the solutions to the transfer gap proposed in recent literature are as follows:

- The JASAC algorithm proposed in [100] makes use of an adversarial agent to learn the control policies that are robust to the transfer gap. The protagonist actor and the adversary actor share a joint critic in the proposed JASAC algorithm that promotes the efficiency and the convergence of the training process, especially for large state-action spaces.
- The SAC algorithm proposed in [94] utilizes a probabilistic ensemble deep neural network (PEDNN) model of the actual distribution network that captures the aleatoric and epistemic uncertainties of voltage. Some of the aleatoric uncertainties of the distribution network are the resistance and reactance of lines that change according to external factors such as humidity and temperature. Epistemic uncertainties account for the uncertainty of the model due to the lack of sufficient data.
- The DDPG algorithm proposed in [92] makes use of a data-driven DNN model that acts as a substitute to the actual electricity distribution network. The constructed DNN model is used to generate the training samples required for the DDPG algorithm without interacting with the actual distribution network.

Satisfaction of network operational constraints is vital in guaranteeing the safety of any proposed RL algorithm. Several different ways that recently proposed RL algorithms implement constraints in the voltage control problem are embedding constraint violations as a penalty term in the formulated reward function, explicitly modelling network constraints as a cost function [98], and through a safety layer in which a sensitivity matrix model is used to predict the change in constrained states over a single time step and correct the actions if required [92].

An RL algorithm that is robust to communication breakdown is more desirable, as it ensures safety in implementation. MARL algorithms are in general more robust to communication breakdown due to their decentralized execution structure. GNN-based policies are identified to be more robust to communication failure and data misalignment than using a conventional dense neural network [75]. The asynchronous learning, sampling and control process proposed in [108] is another solution found in recent literature that guarantees robust communication and minimizes the delay in control actions due to the training process.

Electricity distribution networks frequently encounter fault scenarios and network topology variations. Hence, it is desirable for the RL algorithms used to solve the voltage control problem to be resilient to such network variations. GNN-based policies such as GCN and GAT are commonly used in recent literature to capture topological information in the distribution network. A comparison in the performance of GCN and GAT-based policies for RL algorithms is presented in [77], and the results indicate that GAT-based policies achieve higher performance than GCN due to its attention mechanism that distinguishes important information.

Most of the proposed RL algorithms are simulated on small-scale power systems with a few controllable devices, and it is reasonable to question the scalability of such algorithms. Electricity distribution networks are large-scale systems with multiple controllable devices and numerous uncertainties. Q-learning and DQN algorithms in general are not suited for large-scale systems, as their action space increases exponentially (curse of dimensionality) for every addition of a controllable element, resulting in large policy networks that are

infeasible and difficult to train. The discrete action space of DQN algorithms is another factor that negatively affects their scalability. A wide adoption of SAC algorithms is seen in recent literature, since SAC is more scalable than DQN in addition to its ability to feature both continuous and discrete action spaces. One of such variations of the SAC algorithm that controls devices with discrete action spaces is presented in [98], where the policy network is designed with a device-decoupled structure, such that the network structure only increases linearly with each addition of a controlled device.

The curse of dimensionality is a recurrent problem even in MARL methods, as the input size of the critic network increases with the number of controllable elements. This may result in performance degradation of MARL algorithms when applied to large distribution networks with a high number of control elements. A solution to this scalability issue is proposed in [107], which uses an attention critic that allows intelligent learning of specific information that is most relevant to the rewards. In general, MARL algorithms can be considered to be more scalable than single-agent RL algorithms, as they are less reliant on the communication infrastructure that could introduce delays to the control actions.

8. Conclusions

This paper reviews recently published literature for HC assessment strategies and RL-based coordinated voltage control schemes that enhance the HC of electricity distribution networks. Among the HC quantification methods, QSTS simulation methods are identified as the most accurate approach to evaluate the HC, as it captures the dynamic events of controlled elements. The advent of artificial intelligence in control problems is inevitable and the voltage control problem of power systems is of no exception. However, to the best of the authors' knowledge, no real-world implementation of RL algorithms for coordinated voltage control has been reported yet. This is mainly due to the associated risks and the significantly high start-up cost of such proposals. Most of the literature published on RL algorithms for coordinated voltage control do not undertake an HC assessment to identify the increased HC due to the proposed control algorithm. Undertaking a hosting capacity assessment will provide insights into the techno-economic benefits of the proposed coordinated voltage control schemes and encourage DNSPs and stakeholders to invest more funds in the development of such control schemes. The review of the RL-based algorithms for coordinated voltage control presented in this paper can guide researchers to further develop such RL-based algorithms suitable for practical applications. Future directions for current work include the development of highly accurate HC quantification methods and HC enhancement strategies using RL-based algorithms.

Author Contributions: Conceptualization, J.S., D.R. and A.R.; methodology, J.S. and A.R.; original draft preparation, J.S.; review and editing, A.R. and D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The authors wish to acknowledge the support of Endeavour Energy through the Australian Power Quality and Reliability Centre in providing funding for the resources to complete the above work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Energy Agency. *Renewables 2022 Analysis and Forecast to 2027*; International Energy Agency: Paris, France, 2022. Available online: https://www.iea.org/reports/renewables-2022 (accessed on 10 November 2022).
2. Ding, F.; Mather, B.; Gotseff, P. Technologies to Increase PV Hosting Capacity in Distribution Feeders. In Proceedings of the 2016 IEEE Power and Energy Society General Meeting (PESGM), Boston, MA, USA, 17–21 July 2016. [CrossRef]
3. Rajabi, A.; Elphick, S.; David, J.; Pors, A.; Robinson, D. Innovative approaches for assessing and enhancing the hosting capacity of PV-rich distribution networks: An Australian perspective. *Renew. Sustain. Energy Rev.* **2022**, *161*, 112365. [CrossRef]

4. Chathurangi, D.; Jayatunga, U.; Perera, S.; Agalgaonkar, A.P.; Siyambalapitiya, T. Comparative evaluation of solar PV hosting capacity enhancement using Volt-VAr and Volt-Watt control strategies. *Renew. Energy* **2021**, *177*, 1063–1075. [CrossRef]
5. Liu, J.; Li, Y.; Rehtanz, C.; Cao, Y.; Qiao, X.; Lin, G.; Song, Y.; Sun, C. An OLTC-inverter coordinated voltage regulation method for distribution network with high penetration of PV generations. *Electr. Power Energy Syst.* **2019**, *113*, 991–1001. [CrossRef]
6. Li, C.; Chen, Y.A.; Jin, C.; Sharma, R.; Kleissl, J. Online PV Smart Inverter Coordination using Deep Deterministic Policy Gradient. *Electr. Power Syst. Res.* **2022**, *209*, 107988. [CrossRef]
7. Tomin, N.; Voropai, N.; Kurbatsky, V.; Rehtanz, C. Management of voltage flexibility from inverter-based distributed generation using multi-agent reinforcement learning. *Energies* **2021**, *14*, 8270. [CrossRef]
8. Breker, S.; Rentmeister, J.; Sick, B.; Braun, M. Hosting capacity of low-voltage grids for distributed generation: Classification by means of machine learning techniques. *Appl. Soft Comput. J.* **2018**, *70*, 195–207. [CrossRef]
9. Wu, J.; Yuan, J.; Weng, Y.; Ayyanar, R. Spatial-Temporal Deep Learning for Hosting Capacity Analysis in Distribution Grids. *IEEE Trans. Smart Grid* **2022**, *14*, 354–364. [CrossRef]
10. Mulenga, E.; Bollen, M.H.J.; Etherden, N. A review of hosting capacity quantification methods for photovoltaics in low-voltage distribution grids. *Int. J. Electr. Power Energy Syst.* **2020**, *115*, 105445. [CrossRef]
11. Ismael, S.M.; Abdel Aleem, S.H.E.; Abdelaziz, A.Y.; Zobaa, A.F. State-of-the-art of hosting capacity in modern power systems with distributed generation. *Renew. Energy* **2019**, *130*, 1002–1020. [CrossRef]
12. Chen, X.; Qu, G.; Tang, Y.; Low, S.; Li, N. Reinforcement Learning for Selective Key Applications in Power Systems: Recent Advances and Future Challenges. *IEEE Trans. Smart Grid* **2022**, *13*, 2935–2958. [CrossRef]
13. Wu, M.; Hong, L.; Wang, Y.; Yan, Z.; Chen, Z. Volt-VAR control for distribution networks with high penetration of DGs: An overview. *Electr. J.* **2022**, *35*, 107130. [CrossRef]
14. Yang, T.; Zhao, L.; Li, W.; Zomaya, A.Y. Reinforcement learning in sustainable energy and electric systems: A survey. *Annu. Rev. Control* **2020**, *49*, 145–163. [CrossRef]
15. Cao, D.; Hu, W.; Zhao, J.; Zhang, G.; Zhang, B.; Liu, Z.; Chen, Z.; Blaabjerg, F. Reinforcement Learning and Its Applications in Modern Power and Energy Systems: A Review. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1029–1042. [CrossRef]
16. Kharrazi, A.; Sreeram, V.; Mishra, Y. Assessment techniques of the impact of grid-tied rooftop photovoltaic generation on the power quality of low voltage distribution network—A review. *Renew. Sustain. Energy Rev.* **2019**, *120*, 109643. [CrossRef]
17. Electric Power Research Institute. *EPRI Stochastic Analysis to Determine Feeder Hosting Capacity for Distributed Solar PV*; EPRI Technical Update; Product ID 1026640; Electric Power Research Institute: Washington, DC, USA, 2012; pp. 1–50.
18. Torquato, R.; Salles, D.; Pereira, C.O.; Meira, P.C.M.; Freitas, W. A Comprehensive Assessment of PV Hosting Capacity on Low-Voltage Distribution Systems. *IEEE Trans. Power Deliv.* **2018**, *33*, 1002–1012. [CrossRef]
19. Zubo, R.H.A.; Mokryani, G.; Rajamani, H.S.; Aghaei, J.; Niknam, T.; Pillai, P. Operation and planning of distribution networks with integration of renewable distributed generators considering uncertainties: A review. *Renew. Sustain. Energy Rev.* **2017**, *72*, 1177–1198. [CrossRef]
20. Ebe, F.; Idlbi, B.; Morris, J.; Heilscher, G.; Meier, F. Evaluation of PV hosting capacities of distribution grids with utilisation of solar roof potential analyses. *CIRED Open Access Proc. J.* **2017**, *2017*, 2265–2269. [CrossRef]
21. Ebe, F.; Idlbi, B.; Morris, J.; Heilscher, G.; Meier, F. Evaluation of PV Hosting Capacity of Distribuion Grids Considering a Solar Roof Potential Analysis—Comparison of Different Algorithms. In Proceedings of the 2017 IEEE Manchester PowerTech, Powertech, Manchester, UK, 18–22 June 2017. [CrossRef]
22. Heslop, S.; Macgill, I.; Fletcher, J.; Lewis, S. Method for determining a PV generation limit on low voltage feeders for evenly distributed PV and Load. *Energy Procedia* **2014**, *57*, 207–216. [CrossRef]
23. Carollo, R.; Chaudhary, S.K.; Pillai, J.R. Hosting Capacity of Solar Photovoltaics in Distribution Grids under Different Pricing Schemes. In Proceedings of the 2015 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Brisbane, QLD, Australia, 15–18 November 2015; pp. 5–9. [CrossRef]
24. Heslop, S.; MacGill, I.; Fletcher, J. Maximum PV generation estimation method for residential low voltage feeders. *Sustain. Energy Grids Netw.* **2016**, *7*, 58–69. [CrossRef]
25. Papaioannou, I.T.; Purvins, A. A methodology to calculate maximum generation capacity in low voltage distribution feeders. *Int. J. Electr. Power Energy Syst.* **2014**, *57*, 141–147. [CrossRef]
26. Sexauer, J.M.; Mohagheghi, S. Voltage quality assessment in a distribution system with distributed generation—A probabilistic load flow approach. *IEEE Trans. Power Deliv.* **2013**, *28*, 1652–1662. [CrossRef]
27. Kharrazi, A.; Sreeram, V.; Mishra, Y. Assessment of Voltage Unbalance Due to Single Phase Rooftop Photovoltaic Panels in Residential Low Voltage Distribution Network: A Study on a Real LV Network in Western Australia. In Proceedings of the 2017 Australasian Universities Power Engineering Conference (AUPEC), Melbourne, VIC, Australia, 19–22 November 2018; pp. 1–6. [CrossRef]
28. Vallée, F.; Klonari, V.; Lisiecki, T.; Durieux, O.; Moiny, F.; Lobry, J. Development of a probabilistic tool using Monte Carlo simulation and smart meters measurements for the long term analysis of low voltage distribution grids with photovoltaic generation. *Int. J. Electr. Power Energy Syst.* **2013**, *53*, 468–477. [CrossRef]
29. Bracale, A.; Caramia, P.; Carpinelli, G.; Di Fazio, A.R.; Varilone, P. A bayesian-based approach for a short-term steady-state forecast of a smart grid. *IEEE Trans. Smart Grid* **2013**, *4*, 1760–1771. [CrossRef]

30. Panigrahi, B.K.; Sahu, S.K.; Nandi, R.; Nayak, S. Probablistic Load Flow of a Distributed Generation Connected Power System by Two Point Estimate Method. In Proceedings of the 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Kollam, India, 20–21 April 2017; pp. 1–5.
31. Aien, M.; Fotuhi-Firuzabad, M.; Aminifar, F. Probabilistic load flow in correlated uncertain environment using unscented transformation. *IEEE Trans. Power Syst.* **2012**, *27*, 2233–2241. [CrossRef]
32. Schwippe, J.; Krause, O.; Rehtanz, C. Extension of a Probabilistic Load Flow Calculation Based on an Enhanced Convolution Technique. In Proceedings of the 2009 IEEE PES/IAS Conference on Sustainable Alternative Energy (SAE), Valencia, Spain, 28–30 September 2009; pp. 1–6. [CrossRef]
33. Schellenberg, A.; Rosehart, W.; Aguado, J. Cumulant-based probabilistic optimal power flow (P-OPF) with Gaussian and Gamma distributions. *IEEE Trans. Power Syst.* **2005**, *20*, 773–781. [CrossRef]
34. Chen, W.; Li, X.; Pei, X. Probabilistic Load Flow Calculation Method Based on Polynomial Normal Transformation and Extended Latin hypercube. In Proceedings of the 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2), Changsha, China, 8–10 November 2019; pp. 1369–1374. [CrossRef]
35. Kabir, M.N.; Mishra, Y.; Bansal, R.C. Probabilistic load flow for distribution systems with uncertain PV generation. *Appl. Energy* **2016**, *163*, 343–351. [CrossRef]
36. *IEEE Std 1547.7-2013*; IEEE Guide for Conducting Distribution Impact Studies for Distributed Resource Interconnection. Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2014; pp. 1–137. [CrossRef]
37. Reno, M.J.; Deboever, J.; Mather, B. Motivation and Requirements for Quasi-Static Time Series (QSTS) for Distribution System Analysis. In Proceedings of the 2017 IEEE Power & Energy Society General Meeting, Chicago, IL, USA, 16–20 July 2017; pp. 1–5.
38. Quiroz, J.E.; Reno, M.J.; Broderick, R.J. Time Series Simulation of Voltage Regulation Device Control Modes. In Proceedings of the 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), Tampa, FL, USA, 16–21 June 2013; pp. 1700–1705. [CrossRef]
39. Pecenak, Z.K.; Disfani, V.R.; Reno, M.J.; Kleissl, J. Multiphase Distribution Feeder Reduction. *IEEE Trans. Power Syst.* **2018**, *33*, 1320–1328. [CrossRef]
40. Deboever, J.; Grijalva, S.; Reno, M.J.; Broderick, R.J. Fast Quasi-Static Time-Series (QSTS) for yearlong PV impact studies using vector quantization. *Sol. Energy* **2018**, *159*, 538–547. [CrossRef]
41. López, C.D.; Idlbi, B.; Stetz, T.; Braun, M. Shortening Quasi-Static Time-Series Simulations for Cost-Benefit Analysis of Low Voltage Network Operation with Photovoltaic Feed-In. In Proceedings of the Power and Energy Student Summit (PESS) 2015, Dortmund, Germany, 13–14 January 2015. [CrossRef]
42. Qureshi, M.U.; Grijalva, S.; Reno, M.J.; Deboever, J.; Zhang, X.; Broderick, R.J. A Fast Scalable Quasi-Static Time Series Analysis Method for PV Impact Studies Using Linear Sensitivity Model. *IEEE Trans. Sustain. Energy* **2019**, *10*, 301–310. [CrossRef]
43. Jain, A.K.; Horowitz, K.; Ding, F.; Sedzro, K.S.; Palmintier, B.; Mather, B.; Jain, H. Dynamic hosting capacity analysis for distributed photovoltaic resources—Framework and case study. *Appl. Energy* **2020**, *280*, 115633. [CrossRef]
44. Antoniadou-Plytaria, K.E.; Kouveliotis-Lysikatos, I.N.; Georgilakis, P.S.; Hatziargyriou, N.D. Distributed and Decentralized Voltage Control of Smart Distribution Networks: Models, Methods, and Future Research. *IEEE Trans. Smart Grid* **2017**, *8*, 2999–3008. [CrossRef]
45. Pippi, K.D.; Kryonidis, G.C.; Nousdilis, A.I.; Papadopoulos, T.A. A unified control strategy for voltage regulation and congestion management in active distribution networks. *Electr. Power Syst. Res.* **2022**, *212*, 108648. [CrossRef]
46. Xu, T.; Wade, N.S.; Davidson, E.M.; Taylor, P.C.; McArthur, S.D.J.; Garlick, W.G. Case-based reasoning for coordinated voltage control on distribution networks. *Electr. Power Syst. Res.* **2011**, *81*, 2088–2098. [CrossRef]
47. Jabr, R.A. Linear Decision Rules for Control of Reactive Power by Distributed Photovoltaic Generators. *IEEE Trans. Power Syst.* **2018**, *33*, 2165–2174. [CrossRef]
48. Li, P.; Zhang, C.; Wu, Z.; Xu, Y.; Hu, M.; Dong, Z. Distributed Adaptive Robust Voltage/VAR Control with Network Partition in Active Distribution Networks. *IEEE Trans. Smart Grid* **2020**, *11*, 2245–2256. [CrossRef]
49. Li, J.; Liu, C.; Khodayar, M.E.; Wang, M.H.; Xu, Z.; Zhou, B.; Li, C. Distributed Online VAR Control for Unbalanced Distribution Networks with Photovoltaic Generation. *IEEE Trans. Smart Grid* **2020**, *11*, 4760–4772. [CrossRef]
50. Liu, H.J.; Shi, W.; Zhu, H. Distributed voltage control in distribution networks: Online and robust implementations. *IEEE Trans. Smart Grid* **2018**, *9*, 6106–6117. [CrossRef]
51. Li, Z.; Wu, Q.; Chen, J.; Huang, S.; Shen, F. Double-time-scale distributed voltage control for unbalanced distribution networks based on MPC and ADMM. *Int. J. Electr. Power Energy Syst.* **2023**, *145*, 108665. [CrossRef]
52. Li, P.; Ji, H.; Yu, H.; Zhao, J.; Wang, C.; Song, G.; Wu, J. Combined decentralized and local voltage control strategy of soft open points in active distribution networks. *Appl. Energy* **2019**, *241*, 613–624. [CrossRef]
53. Farina, M.; Guagliardi, A.; Mariani, F.; Sandroni, C.; Scattolini, R. Model predictive control of voltage profiles in MV networks with distributed generation. *Control Eng. Pract.* **2015**, *34*, 18–29. [CrossRef]
54. Papadimitrakis, M.; Kapnopoulos, A.; Tsavartzidis, S.; Alexandridis, A. A cooperative PSO algorithm for Volt-VAR optimization in smart distribution grids. *Electr. Power Syst. Res.* **2022**, *212*, 108618. [CrossRef]
55. Nayeripour, M.; Fallahzadeh-Abarghouei, H.; Waffenschmidt, E.; Hasanvand, S. Coordinated online voltage management of distributed generation using network partitioning. *Electr. Power Syst. Res.* **2016**, *141*, 202–209. [CrossRef]
56. Zhao, B.; Xu, Z.; Xu, C.; Wang, C.; Lin, F. Network Partition-Based Zonal Voltage Control for Distribution Networks With Distributed PV Systems. *IEEE Trans. Smart Grid* **2018**, *9*, 4087–4098. [CrossRef]

57. Zhang, Y.; Wang, X.; Wang, J.; Zhang, Y. Deep Reinforcement Learning Based Volt-VAR Optimization in Smart Distribution Systems. *IEEE Trans. Smart Grid* **2021**, *12*, 361–371. [CrossRef]
58. El Helou, R.; Kalathil, D.; Xie, L. Fully Decentralized Reinforcement Learning-Based Control of Photovoltaics in Distribution Grids for Joint Provision of Real and Reactive Power. *IEEE Open Access J. Power Energy* **2021**, *8*, 175–185. [CrossRef]
59. Liu, H.; Wu, W. Federated Reinforcement Learning for Decentralized Voltage Control in Distribution Networks. *IEEE Trans. Smart Grid* **2022**, *13*, 3840–3843. [CrossRef]
60. Sutton, R.S.; Barto, A.G. Adaptive Computation and Machine Learning. In *Reinforcement Learning: An Introduction, 2nd ed*; The MIT Press: Cambridge, MA, USA, 2018; ISBN 9780262039246.
61. Gupta, N.; Chandwani, V. Artificial Neural Networks as Universal Function Approximators. *Int. J. Emerg. Trends Eng. Dev.* **2012**, *4*, 456–464.
62. Ghayoumi, M. *Deep Learning in Practice, 1st ed*; CRC Press: Boca Raton, FL, USA, 2022; ISBN 9781003025818.
63. Zhang, Y.; Chen, D.; Ye, C. *Toward Deep Neural Networks: WASD Neuronet Models, Algorithms, and Applications*; Chapman & Hall/CRC Artificial Intelligence and Robotics Series; CRC Press: Boca Raton, FL, USA, 2019; ISBN 0-429-76098-1.
64. Lotfi, A.; Pirnia, M. Constraint-guided Deep Neural Network for solving Optimal Power Flow. *Electr. Power Syst. Res.* **2022**, *211*, 108353. [CrossRef]
65. Sun, R. Optimization for Deep Learning: Theory and Algorithms. *arXiv* **2019**, arXiv:1912.08957.
66. Shi, Z.; Yao, W.; Zeng, L.; Wen, J.; Fang, J.; Ai, X.; Wen, J. Convolutional neural network-based power system transient stability assessment and instability mode prediction. *Appl. Energy* **2020**, *263*, 114586. [CrossRef]
67. Zou, M.; Zhao, Y.; Yan, D.; Tang, X.; Duan, P.; Liu, S. Double convolutional neural network for fault identification of power distribution network. *Electr. Power Syst. Res.* **2022**, *210*, 108085. [CrossRef]
68. Wang, S.; Chen, H. A novel deep learning method for the classification of power quality disturbances using deep convolutional neural network. *Appl. Energy* **2019**, *235*, 1126–1140. [CrossRef]
69. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
70. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
71. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [CrossRef]
72. Dong, H.; Zhu, J.; Li, S.; Wu, W.; Zhu, H.; Fan, J. Short-term residential household reactive power forecasting considering active power demand via deep Transformer sequence-to-sequence networks. *Appl. Energy* **2023**, *329*, 120281. [CrossRef]
73. Mellit, A.; Pavan, A.M.; Lughi, V. Deep learning neural networks for short-term photovoltaic power forecasting. *Renew. Energy* **2021**, *172*, 276–288. [CrossRef]
74. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81. [CrossRef]
75. Lee, X.Y.; Sarkar, S.; Wang, Y. A graph policy network approach for Volt-Var Control in power distribution systems. *Appl. Energy* **2022**, *323*, 119530. [CrossRef]
76. Hossain, R.R.; Huang, Q.; Huang, R. Graph convolutional network-based topology embedded deep reinforcement learning for voltage stability control. *IEEE Trans. Power Syst.* **2021**, *36*, 4848–4851. [CrossRef]
77. Xing, Q.; Chen, Z.; Zhang, T.; Li, X.; Sun, K.H. Real-time optimal scheduling for active distribution networks: A graph reinforcement learning method. *Int. J. Electr. Power Energy Syst.* **2023**, *145*, 108637. [CrossRef]
78. Zhang, H.; Yu, T. *Taxonomy of Reinforcement Learning Algorithms BT—Deep Reinforcement Learning: Fundamentals, Research and Applications*; Dong, H., Ding, Z., Zhang, S., Eds.; Springer: Singapore, 2020; pp. 125–133. ISBN 978-981-15-4095-0.
79. Vlachogiannis, J.G.; Hatziargyriou, N.D. Reinforcement learning for reactive power control. *IEEE Trans. Power Syst.* **2004**, *19*, 1317–1325. [CrossRef]
80. Van Hasselt, H.; Guez, A.; Silver, D. Deep Reinforcement Learning with Double Q-Learning. *AAAI Conf. Artif. Intell.* **2016**, *30*, 2094–2100. [CrossRef]
81. Wang, Z.; Schaul, T.; Hessel, M.; Van Hasselt, H.; Lanctot, M.; De Frcitas, N. Dueling Network Architectures for Deep Reinforcement Learning. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 2939–2947.
82. Fortunato, M.; Azar, M.G.; Piot, B.; Menick, J.; Hessel, M.; Osband, I.; Graves, A.; Mnih, V.; Munos, R.; Hassabis, D.; et al. Noisy Networks for Exploration. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–21.
83. Bellemare, M.G.; Dabney, W.; Munos, R. A Distributional Perspective on Reinforcement Learning. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 693–711.
84. Yang, Q.; Wang, G.; Sadeghi, A.; Giannakis, G.B.; Sun, J. Two-Timescale Voltage Control in Distribution Grids Using Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2020**, *11*, 2313–2323. [CrossRef]
85. Schulman, J.; Levine, S.; Moritz, P.; Jordan, M.I.; Abbeel, P. Trust Region Policy Optimization. *arXiv* **2015**, arXiv:1502.05477.
86. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.

87. Sutton, R.S.; McAllester, D.; Singh, S.; Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst.* **2000**, *32*, 1057–1063.
88. Cao, D.; Hu, W.; Xu, X.; Wu, Q.; Huang, Q.; Chen, Z.; Blaabjerg, F. Deep Reinforcement Learning Based Approach for Optimal Power Flow of Distribution Networks Embedded with Renewable Energy and Storage Devices. *J. Mod. Power Syst. Clean Energy* **2021**, *9*, 1101–1110. [CrossRef]
89. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic policy gradient algorithms. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing China, 21–26 June 2014; Volume 1, pp. 605–619.
90. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. In Proceedings of the 4th International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2–4 May 2016.
91. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Yu, N.; Huang, Q.; Chen, Z. Model-free voltage control of active distribution system with PVs using surrogate model-based deep reinforcement learning. *Appl. Energy* **2022**, *306*, 117982. [CrossRef]
92. Kou, P.; Liang, D.; Wang, C.; Wu, Z.; Gao, L. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. *Appl. Energy* **2020**, *264*, 114772. [CrossRef]
93. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *arXiv* **2018**, arXiv:1801.01290.
94. Liu, Q.; Guo, Y.; Deng, L.; Tang, W.; Sun, H.; Huang, W. Robust Offline Deep Reinforcement Learning for Volt-Var Control in Active Distribution Networks. In Proceedings of the 2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2), Taiyuan, China, 22–24 October 2021; pp. 442–448. [CrossRef]
95. Liu, H.; Wu, W.; Wang, Y. Bi-level Off-policy Reinforcement Learning for Two-Timescale Volt/VAR Control in Active Distribution Networks. *IEEE Trans. Power Syst.* **2022**, *38*, 385–395. [CrossRef]
96. Yang, Q.; Simão, T.D.; Tindemans, S.; Spaan, M.T.J. WCSAC: Worst-Case Soft Actor Critic for Safety-Constrained Reinforcement Learning. *AAAI Conf. Artif. Intell.* **2021**, *35*, 10639–10646. [CrossRef]
97. Ha, S.; Xu, P.; Tan, Z.; Levine, S.; Tan, J. Learning to Walk in the Real World with Minimal Human Effort. *arXiv* **2020**, arXiv:2002.08550.
98. Wang, W.; Yu, N.; Gao, Y.; Shi, J. Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems. *IEEE Trans. Smart Grid* **2020**, *11*, 3008–3018. [CrossRef]
99. Wang, W.; Yu, N.; Shi, J.; Gao, Y. Volt-VAR Control in Power Distribution Systems with Deep Reinforcement Learning. In Proceedings of the 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Beijing, China, 21–23 October 2019. [CrossRef]
100. Liu, H.; Wu, W. Two-Stage Deep Reinforcement Learning for Inverter-Based Volt-VAR Control in Active Distribution Networks. *IEEE Trans. Smart Grid* **2021**, *12*, 2037–2047. [CrossRef]
101. Gao, Y.; Wang, W.; Yu, N. Consensus Multi-Agent Reinforcement Learning for Volt-VAR Control in Power Distribution Networks. *IEEE Trans. Smart Grid* **2021**, *12*, 3594–3604. [CrossRef]
102. Xu, Y.; Zhang, W.; Liu, W.; Ferrese, F. Multiagent-Based Reinforcement Learning for Optimal Reactive Power Dispatch. *IEEE Trans. Syst. Man, Cybern. Part C Appl. Rev.* **2012**, *42*, 1742–1751. [CrossRef]
103. Nguyen, T.T.; Nguyen, N.D.; Nahavandi, S. Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications. *IEEE Trans. Cybern.* **2020**, *50*, 3826–3839. [CrossRef]
104. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *arXiv* **2017**, arXiv:1706.02275.
105. Sun, X.; Qiu, J. Two-Stage Volt/Var Control in Active Distribution Networks with Multi-Agent Deep Reinforcement Learning Method. *IEEE Trans. Smart Grid* **2021**, *12*, 2903–2912. [CrossRef]
106. Wang, S.; Duan, J.; Shi, D.; Xu, C.; Li, H.; Diao, R.; Wang, Z. A Data-Driven Multi-Agent Autonomous Voltage Control Framework Using Deep Reinforcement Learning. *IEEE Trans. Power Syst.* **2020**, *35*, 4644–4654. [CrossRef]
107. Cao, D.; Hu, W.; Zhao, J.; Huang, Q.; Chen, Z.; Blaabjerg, F. A Multi-Agent deep reinforcement learning based voltage regulation using coordinated pv inverters. *IEEE Trans. Power Syst.* **2020**, *35*, 4120–4123. [CrossRef]
108. Liu, H.; Wu, W. Online Multi-Agent Reinforcement Learning for Decentralized Inverter-Based Volt-VAR Control. *IEEE Trans. Smart Grid* **2021**, *12*, 2980–2990. [CrossRef]
109. Li, C.; Jin, C.; Sharma, R. Coordination of PV Smart Inverters using Deep Reinforcement Learning for Grid Voltage Regulation. In Proceedings of the 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1930–1937. [CrossRef]
110. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Huang, Q.; Chen, Z.; Blaabjerg, F. Data-Driven Multi-Agent Deep Reinforcement Learning for Distribution System Decentralized Voltage Control with High Penetration of PVs. *IEEE Trans. Smart Grid* **2021**, *12*, 4137–4150. [CrossRef]
111. Zhang, X.; Liu, Y.; Duan, J.; Qiu, G.; Liu, T.; Liu, J. DDPG-Based Multi-Agent Framework for SVC Tuning in Urban Power Grid with Renewable Energy Resources. *IEEE Trans. Power Syst.* **2021**, *36*, 5465–5475. [CrossRef]
112. Cao, D.; Zhao, J.; Hu, W.; Yu, N.; Ding, F.; Huang, Q.; Chen, Z. Deep Reinforcement Learning Enabled Physical-Model-Free Two-Timescale Voltage Control Method for Active Distribution Systems. *IEEE Trans. Smart Grid* **2022**, *13*, 149–165. [CrossRef]

113. Li, H.; He, H. Learning to Operate Distribution Networks With Safe Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2022**, *13*, 1860–1872. [CrossRef]
114. Hu, D.; Peng, Y.; Yang, J.; Deng, Q.; Cai, T. Deep Reinforcement Learning Based Coordinated Voltage Control in Smart Distribution Network. In Proceedings of the 2021 International Conference on Power System Technology (POWERCON), Haikou, China, 8–9 December 2021; pp. 1030–1034. [CrossRef]