IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 May 2024, accepted 16 May 2024, date of publication 20 May 2024, date of current version 28 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3403035

SURVEY

Data-Driven Volt/VAR Optimization for Modern Distribution Networks: A Review

SARAH ALLAHMORADI[1], (Member, IEEE), SHAHABODIN AFRASIABI[1], (Member, IEEE), XIAODONG LIANG[1], (Senior Member, IEEE), JUNBO ZHAO[2], (Senior Member, IEEE), AND MOHAMMAD SHAHIDEHPOUR[3], (Life Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada
[2]Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269, USA
[3]Robert W. Galvin Center for Electricity Innovation, Illinois Institute of Technology, Chicago, IL 60616, USA

Corresponding author: Xiaodong Liang (xil659@mail.usask.ca)

This work was supported in part by the University of Saskatchewan, Saskatoon, SK, Canada.

ABSTRACT The Volt/Var optimization (VVO) enables advanced control strategy development for voltage regulation. With the recent advancement of data-driven approaches and communication infrastructure, realtime decision-making through VVO can effectively address distributed energy resources (DERs) uncertainties without relying on models and topologies of distribution networks. In this paper, a comprehensive review on data-driven VVO in distribution networks is presented, focusing on statistics and machine learning (supervised/unsupervised, ensemble, and reinforcement learning (RL)). State-of-the-art monitoring devices essential in data-driven VVO frameworks are firstly discussed. How data-driven structures serve as primary or supplementary tools in VVO frameworks is then detailed. Since RL is increasingly used, RL-based algorithms (value-based, policy-based, actor-critic-based, and graph-based algorithms) are reviewed. Decision-making processes for RL-based VVO frameworks, such as the Markov decision process (MDP), Markov game, constrained Markov decision process, constrained Marko game, and adversarial Markov decision process, are also surveyed. Future research directions in this area are recommended in the paper.

INDEX TERMS Data-driven decision-making, distribution networks, supervised, unsupervised, ensemble learning, reinforcement learning, renewable energy resources, Volt/Var optimization.

I. INTRODUCTION

Distributed energy resources (DERs) include renewable energy sources (RESs) (solar, wind, and hydropower, etc.), small-scale fossil fuel-based generation technologies, energy storage systems (ESSs), demand response (DR) programs, and electric vehicles (EVs). Their increasing penetration in modern distribution networks (MDNs) creates new control and operational challenges [1]. According to the world energy transition roadmap from the international renewable energy agency (IRENA), the proportion of RESs is projected to increase from 14% in 2019 to 40% by 2030 [2]. Distribution networks with high penetration of RESs may experience voltage instability issues due to power fluctuations of RESs [3]. Voltage variations may significantly affect the system's power quality and reliability, voltage-dependent loads performance, and cause excessive operations of voltage regulation devices. Volt/Var optimization (VVO) is a vital tool to optimize voltage-regulating device operations, reduce voltage deviations, and power losses.

Control architectures for VVO can be broadly classified into model-based [4], [5], [6], [7], [8], [9] and data-driven approaches. Effective model-based approaches rely heavily on accurate knowledge of the grid topologies and parameters [10], which are hard to obtain for real-time operations, as the system topology can be highly variable and complex due to bidirectional power flow among multiple DERs. Thus, the model-based approaches are computationally demanding to obtain accurate predetermined mathematical solutions for the current state of an MDN.

The associate editor coordinating the review of this manuscript and approving it for publication was Rajesh Kumar.

To address these challenges, data-driven approaches are developed to handle complexities and uncertainties of MDNs [11]. The data-driven model is independent of the model of distribution networks and overcomes issues associated with model-based approaches [12], [13]. Although data-driven approaches have been applied in various domains of power systems [14], [15], [16], the data-driven VVO research in MDNs is largely lacking.

Regarding VVO in modern power distribution systems, Ref. [4] focuses on centralized and decentralized control strategies, and examines both traditional devices, such as on-load tap changers and capacitors, as well as newer distributed generation technologies. Ref. [17] introduces deep reinforcement learning to VVO, which can adapt to dynamics of power systems without the detailed network models. Ref. [18] discusses the integration of large-scale wind farms, and highlights challenges and adaptive strategies necessary for voltage stability and power quality. Ref. [19] surveys classical and heuristic optimization methods in VVO, and offers a critical comparison of their effectiveness across various configurations. Ref. [20] reviews general VVO techniques, and uses machine learning to enhance real-time decisionmaking. This paper aims to address limitations in existing reviews in the literature, where traditional model-based strategies are main focus, and how machine learning is integrated with VVO frameworks to enhance grid resilience and efficiency has not been fully explored.

To fill in this research gap, a comprehensive review on data-driven VVO in distribution networks is conducted in this paper. The main contributions of this review include:

- Provide deeper insights into VVO implementations with cutting-edge monitoring and measurement systems, and compare VVO with Volt/Var control (VVC) in MDNs.
- Review data-driven approaches as primary or supplementary tools for VVO, including statistics and machine learning.
- Summarize key aspects of various reinforcement learning (RL)-based VVO algorithms (value-based, policy-based, actor-critic-based, and graph-based algorithms).
- Review decision-making processes for RL-based VVO frameworks (the Markov decision process (MDP), Markov game (MG), constrained Markov decision process (CMDP), constrained Markov game (CMG), and adversarial Markov decision process (AMDP)).

The paper is organized as follows: In Section II, methods and material to conduct this review is introduced; in Section III, VVO is introduced with advanced monitoring and control systems; in Section IV, different model-free decision-making VVO frameworks are introduced, including statistics, and machine learning methods; in Section V, RLbased decision-making frameworks for VVO are introduced; RL algorithms are summarized in Section VI; in Section VII, RL-based VVO frameworks in MDNs are reviewed; Section VIII recommends future research directions in this area; Section IX concludes the paper. To ensure a thorough understanding, discussions on RL-based decision frameworks, algorithms, and their applications are arranged into Sections V, VI and VII, respectively. Each section is dedicated to exploring its specific aspect, foundational theories, detailed algorithmic approaches, and practical applications, thereby providing a clear, step-by-step progression from theoretical concepts to real-world implementations.

II. METHODS AND MATERIALS

This review systematically analyzes the current body of knowledge on data-driven VVO strategies in MDNs. We aim to ensure a comprehensive coverage and rigorous analysis of the literature to identify research gaps, current technologies, and future research directions in this evolving field.

A. RESEARCH QUESTIONS

The research questions guiding this review are:

1) What are the current data-driven approaches utilized in VVO for MDNs?
2) How can these data-driven approaches improve the system performance compared to traditional modelbased strategies?
3) What are challenges and limitations of existing datadriven VVO strategies?
4) What are the future research directions?

B. SEARCH STRATEGY

A systematic literature search was conducted across several databases, including IEEE Xplore, ScienceDirect, and Scopus. Keywords used in the search included ''datadriven VVO,'' ''Volt/Var optimization,'' ''modern distribution networks,'' ''machine learning in power systems,'' and ''reinforcement learning in VVO''. These keywords were combined using Boolean operators to ensure a broad retrieval of relevant studies.

C. SCREENING AND SELECTION CRITERIA

Papers published in peer-reviewed journals and conference proceedings, and studies that specifically discussed datadriven approaches used in VVO are included.

D. DATA EXTRACTION AND SYNTHESIS

Selected papers were subjected to a data extraction process, where key information was categorized based on the datadriven approach employed, specific applications within VVO, benefits and limitations observed, and the geographical focus of the study. A thematic analysis was then conducted to synthesize findings across the selected studies, identifying common themes, trends, and discrepancies in the data-driven VVO literature.

III. VOLT/VAR OPTIMIZATION IN MDNS

In this section, VVC and VVO concepts, advanced control technologies, new monitoring and measurement systems, and improved communication and control infrastructures for VVO in MDNs are introduced.

A. CONCEPTS OF VVC AND VVO

Although VVC and VVO are used interchangeably in the literature, they are two different techniques for voltage and reactive power management to maintain stability, improve power quality, and reduce power losses in a distribution grid. As a traditional local control approach, VVC relies on predetermined settings and rules for voltage control devices, while maintaining the voltage limits of [0.95, 1.05] p.u. according to ANSI C84.1 [21] and CAN 3-C235-83 standards [22], and controlling the reactive power flow.

VVC can be implemented through the legacy voltage regulation devices, such as on-load tap changers (OLTCs), voltage regulators (VR), and capacitor banks (CBs):

- OLTCs via the tap ratio setting control the voltage. The OLTC adjusts the secondary transformer voltage between V_l^{OLTC} and V_u^{OLTC}, and lower and upper voltage bounds of the OLTC, respectively. These bounds are adjusted $V_{set} \pm 0.05\ V_{DB}$, where V_{DB} is the dead band designed to reduce oscillations.
- VR can modify the feeder voltage within a $\pm 10\%$ range.
- CBs regulate Var and voltage by injecting reactive power as follows:

$$Q_c = V_c^2 Q_c^n \tag{1}$$

where Q_C is reactive power injected by a capacitor bank, V_C and Q_c^n are the voltage across the capacitor and its rated reactive power, respectively. For a line in MDNs defined by the resistance R_l and the reactance X_l, the resulting voltage can be described by

$$\Delta V \approx \frac{[R_l P_l + X_l\,(Q_l - Q_c]}{V} \tag{2}$$

Using VVC, the voltage and reactive power regulators are controlled independently by local control systems based on local measurements without needing a communication infrastructure. However, VVC does not provide optimization.

To provide optimization for voltage-regulating devices, model-based optimization strategies have been developed. The common mathematical model for VVO is expressed by

$$min\ F = \sum_{i=1}^{n} |V_i - 1| + C_e P_{loss} \tag{3}$$

$$s.t\ P_{Gi} - P_{Li} - V_i \sum_{j=1}^{n} V_j \left(G_{ij} cos\sigma_{ij} + B_{ij} sin\sigma_{ij}\right) = 0 \tag{4}$$

$$Q_{Gi} + Q_{Ci} - Q_{Li} - V_i \sum_{j=1}^{n} V_j \left(G_{ij} sin\sigma_{ij} - B_{ij} cos\sigma_{ij}\right) = 0 \tag{5}$$

$$\underline{Q}_{ci} \le Q_{ci} \le \bar{Q}_{ci} \tag{6}$$

where F is the objective function, n is the total number of nodes, V_i is the voltage at node i. C_e and P_{loss} are the electricity price and power losses, respectively. P_{Gi} and Q_{Gi} are active and reactive power of generator i, respectively. P_{Li} and Q_{Li} are active and reactive power of the load demand at node i, respectively. Q_{Ci} is reactive power compensated by regulating devices at node i. G_{ij} and B_{ij} are conductance and susceptance of line ij, respectively. σ_{ij} is the phase angle difference between head and tail nodes. $\underline{Q}_{Ei}$ and $\bar{Q}_{Ei}$ are lower and upper regulation limits for reactive power compensated from voltage regulating devices connected to node i, respectively.

It should be noted that the objective functions utilized in VVO are not limited to minimizing voltage deviations and power losses. Additional objectives include reduction in distribution planning expansion costs, the consumer energy usage, and operating and maintenance expenses. These objectives have been extensively employed in prior research to assess the efficacy and cost-efficiency of VVO strategies [19], [23].

VVO is more flexible and adaptable than VVC due to sophisticated communication infrastructures and data analytics, and thus, can continuously adapt to changing grid conditions, effectively incorporate DERs, and optimize operations of voltage and reactive power regulation devices.

B. CONTROL DEVICES

Conventionally, VVO uses utility-owned or slow-response control devices (SRCDs) to manage reactive power and voltage levels at the end of a feeder and compensate for load variations in distribution networks. Traditional voltage regulation devices, such as OLTCs, VRs, and CBs, are electromechanical control devices installed on the primary feeder, and can make informed decisions using local voltage and current measurements at various loading levels [24]. To maintain reliability and stability of power grids, utilityowned devices are strategically installed on feeders that experience a high loading level and at the point of connection with the upstream grid. For example, OLTCs and VRs are typically located at a substation and along a distribution line, respectively. An OLTC enables tap changes while the transformer remains on load, while a VR changes the tap position during a brief interruption. Capacitor banks installed on feeders or substations are usually dispatched to regulate reactive power at day-ahead operations to ensure that the voltage remains within an acceptable range. However, these control devices have limited life cycles, slow control speed, and frequent switching due to high penetrations of DERs, and are not suitable for real-time operations.

Recently, fast-response control devices (FRCDs), such as static var compensators (SVCs), static synchronous compensators (STATCOMs), and smart inverters (SIs), have been progressively used for VVO in real-time operations as recommended in IEEE standard 1547.8 [25]. FRCDs can operate within milliseconds and mitigate rapid voltage fluctuations caused by DERs, so they are more flexible and adaptable than traditional SRCDs [26]. FRCDs can also be easily integrated into an existing system and reprogrammed as needed under changing operating conditions. They can be placed at substations or distribution feeders to maintain the voltage profile and stabilize the grid. To efficiently manage DERs output and enhance power quality, SIs can be strategically placed at the interconnection point between DERs and a distribution

network with reactive power support while reducing power losses and voltage fluctuations [24]. Table 1 provides a comparison of SRCDs and FRCDs.

TABLE 1. Comparison of slow and fast response control devices in VVO.

	SRCDs	FRCDs
Control devices	OLTCs, CBs, VRs	SIs, SVCs, STATCOMs
Operation timescale	Slow	Fast
Control variable	Discrete	Continues
Penetration	Low	High
Installation cost	Low	High
Efficiency	Lower	Higher
Implementation	Simple	Complex
Operation/ maintenance cost	Lower	Higher

DR programs, conventional distributed generation (DG) units, ESSs, and smart transformers (STs) are valuable resources in MDNs. A DR program can balance the supply and demand, allow flexible real-time adjustments to load consumptions based on the system state (for example, solar and wind power fluctuations), reduce the peak load, limit the need for additional infrastructure investments, and reduce the strain on existing grid components, and thus, it can improve voltage regulation and load balancing by actively managing the demand side of a grid. Conventional DGs, such as diesel generators, micro-turbines, and combined heat and power (CHP) systems, provide localized power generation to balance local load and maintain a stable voltage level. With multiple points of power generation across the network, DGs can reduce power losses and improve the grid's resilience. The flexibility of DGs in terms of size and technology allows them to be deployed strategically for VVO. ESs can store excess power during a low-demand period and discharge it during a high-demand period. STs, including solid-state transformers (SSTs) and hybrid distribution transformers (HDTs), are power electronics-based devices, capable of two-way communication and real-time diagnostics to adapt to grid conditions swiftly. Power electronic converters in STs provide reactive power compensation on the mediumvoltage side, a key asset in handling intermittent RESs. Digital control of STs is crucial for implementing VVO. All control devices can be strategically placed, a schematic of control devices for VVO in a MDN is shown in Fig. 1.

C. ADVANCED MONITORING AND MEASUREMENT SYSTEMS

Cutting-edge monitoring and measurement systems ensure improved observability of a distribution network, which VVO strategies rely upon (Fig. 2). At low- and mediumvoltage levels, commonly used monitoring systems include the supervisory control and data acquisition (SCADA), micro phasor measurement units (μ PMUs), power quality monitors (PQM), smart meters (SMs), and intelligent sensors. They offer precise measurements, rapid communication, and remote data storage. Using these devices, real-time data (the voltage, current, frequency, power factor, active and

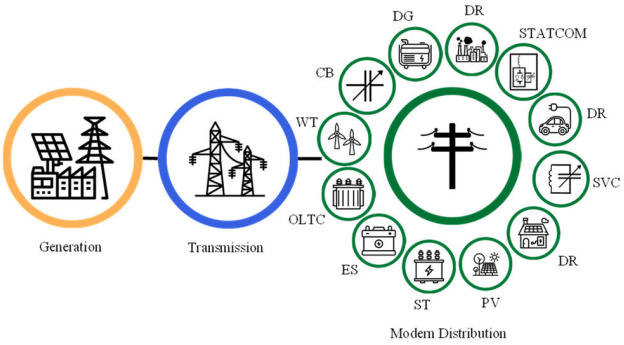


FIGURE 1. Schematic of control devices and technologies for VVO in MDNs.

reactive power) across distribution networks can be measured to support VVO strategies and facilitate DERs integration in MDNs through advanced data analytics, and wired/wireless communication infrastructures (power lines, fiber optics, the wireless radio frequency, cellular networks, and the satellite communication) [27], [28].

D. ADVANCED CONTROL STRATEGIES

Traditional VVC strategies comprise standalone, on-site voltage regulators, and rule-based control approaches, where the controller operates based on rules and historical or online measurement data. Although low-cost and communicationless, traditional VVC strategies lack optimality and coordination among voltage-regulating devices [24]. For example, VVC operates independently without considering impacts of neighboring devices, which may lead to conflict control actions and decreased system efficiency [29].

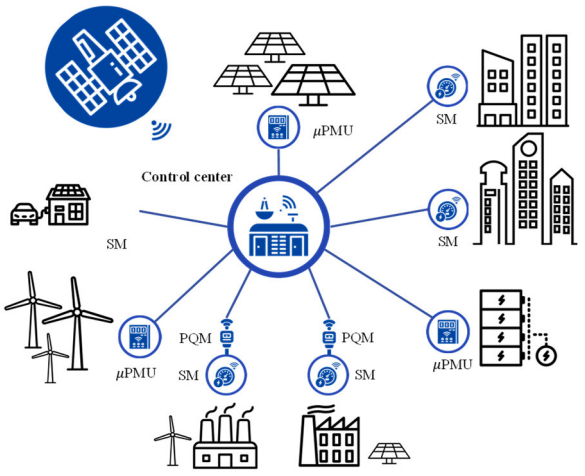


FIGURE 2. Advanced monitoring and measurement systems in a MDN.

In Fig. 3, model-based control strategies for VVO are categorized into centralized, distributed, and decentralized based on communication structures [6]. Centralized control relies on a central controller to make decisions for the entire system

based on optimal power flow, and a fast and reliable communication system to receive measurements from devices, such as smart meters or remote terminal units.

However, due to the massive information exchange and computational burden, centralized control is less effective in local conditions for real-time voltage control; the system is vulnerable to single-point failures [26], and has customer information privacy concerns. Distributed control makes decisions by individual control devices based on optimization of local measurement data [30], [31], which still requires the information exchange with neighboring units. However, each device is responsible for managing the voltage and reactive power, and cannot be remotely dispatched. To reduce communication reliance and offer faster control actions, decentralized control has emerged as a promising technique between centralized and distributed control [32], where the system is divided into small zones, each zone has its own controller [33], and each controller is responsible for managing the voltage and reactive power within its zone using local data [34]. Decentralized control requires less computation and information exchange than centralized control, and is more reliable than distributed control in large distribution grids [9].

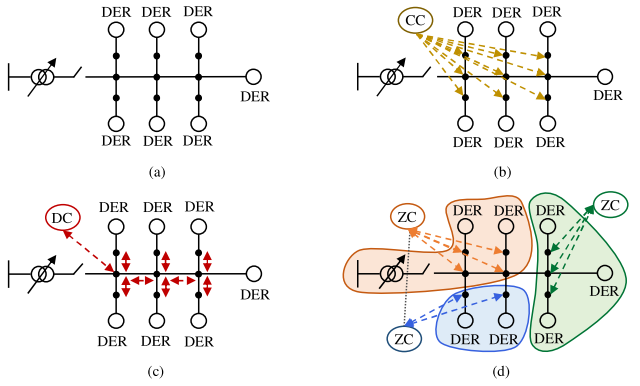


FIGURE 3. Schematic of control strategies for VVO in a MDN based on communication structures: (a) rule-based control without communication, (b) centralized control (CC: Central coordinator), (c) distributed control (DC: Distributed coordinator), (d) decentralized control (ZC: Zone coordinator) [9].

E. MODEL-BASED AND DATA-DRIVEN VVO

In practical implementation, model-based VVO has increasingly been adopted by utilities. These strategies utilize dynamic modeling of MDNs using Geographic information systems (GIS) for monitoring the connectivity of a distribution network. This allows utilities to achieve detailed network and customer load modeling and manage the vast complexity of the distribution grid more effectively. The well-known method within this framework is conservation voltage reduction (CVR), which aims for a reduction of 2-4% in demand by optimizing voltage levels across the network [23]. To solve the VVO through the model-based approaches, mathematical optimization algorithms are applied by utilities, which are subject to millions of nonlinear equality and inequality constraints. This complexity arises from the need to accurately model the distribution system, considering a large number of variables and states that reflect the real-world behavior of the grid. The constraints and variables include but are not limited to, voltage levels, power flows, and operational limits of grid components, such as transformers and CBs.

Broadly, model-based VVO approaches can be classified into two primary categories: classical and heuristic, as depicted in Fig. 4. The classical category covers various methods, including first-order gradient-based, second-order gradient-based, quadratic programming, linear programming, interior-point methods, and mixed-integer programming. First-order gradient-based approaches are iterative optimization techniques that optimize a differentiable nonlinear function through a sequence of decision vectors, utilizing the first-order derivatives of objective functions. Conversely, second-order gradientbased approaches enhance the optimization precision by incorporating second-order derivatives, providing a refined approximation of objective functions. Quadratic programming specializes in optimizing quadratic objective functions subject to linear constraints. In VVO studies employing quadratic programming, sequential quadratic programming techniques are often used, iteratively creating a quadratic approximation of the objective functions alongside a linear approximation of the constraints. Linear programming-based VVO models are employed when objective functions and constraints are linear, considering only continuous decision variables within the VVO framework. The inclusion of discrete variables, representative of SRCDs, necessitates a shift towards mixedinteger programming formulations. These formulations integrate integrality constraints with both continuous and discrete decision variables, accommodating the comprehensive nature of VVO tasks. On the other hand, Heuristic models are search-based optimization techniques that primarily rely on the strategic orientation of decision variables through specific algorithms to expedite convergence towards optimal solutions. These models are characterized by their ability to navigate complex solution spaces efficiently, often reaching satisfactory solutions more swiftly than conventional methods. However, it is important to note that heuristic optimization models do not guarantee the identification of global optimum. The inherent nature of heuristic approaches means they seek to balance computational efficiency with solution quality, making them particularly suitable for large-scale or complex optimization problems where exact methods are computationally infeasible [19], [24], [35].

Given the detailed and comprehensive nature of these models, VVO emerges as a challenging task, particularly considering the large number of states involved in the practical implementation of VVO strategies. These states are essential for capturing the dynamic behavior of the grid, including changes in load demand, generation patterns, and the operational status of grid components. The complexity of managing these states, combined with the computational

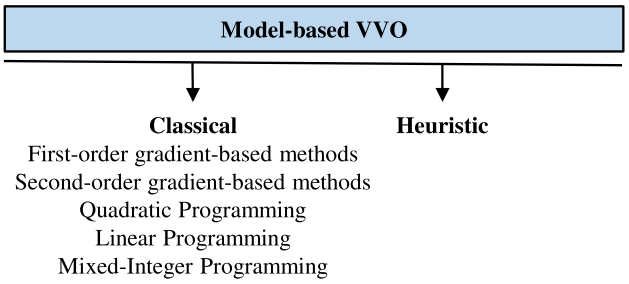


FIGURE 4. The model-based VVO categories.

intensity of solving the optimization problem, underscores the challenges utilities face in implementing effective VVO solutions. Moreover, the model-based VVO relies on an accurate physical model of distribution networks, and its decisions are influenced by many parameters, such as line reactance, line resistance, and transformer tap settings. However, accurately modeling a MDN with a large number of nodes, lines, and buses is an extremely complex task. On the other hand, data-driven VVO is more practical for MDNs due to ample data available and has significantly faster computational speed than model-based VVO without requiring iterative algorithms. Also, the main advantage of recently developed data-driven models is their capability to perform without the requirement of advanced measurement devices. These modern data-driven approaches, such as generative networks, can even perform effectively with small-scale datasets. However, the model-based approaches require precise network parameters and measurements obtained from advanced measurement infrastructures. For adaptability, the model-based VVO's updates are less frequent, and typically synchronized with significant grid modifications and recalibration procedures; while datadriven VVO can dynamically respond to real-time data to realize voltage and reactive power management in the power grid. With the recent advancement of monitoring systems and artificial intelligence techniques, data-driven strategies, particularly RL have gained significant attention. By solving VVO through RL algorithms, a MDN is considered as a black box without requiring the network's topologies and parameters, and the agent makes decisions based on state observations [11]. A comparison of model-based and datadriven VVO is shown in Table 2.

TABLE 2. Comparison of model-based and data-driven VVO.

Criteria	Model-based VVO	Data-Driven VVO
Performance Time	High	Low
Flexibility	Less adaptive	Adaptability depends on the training process
Data Requirement	Needs accurate grid parameters and real-time measurements	A large-scale dataset for training
Advanced measurement device requirement	Yes	No
Update frequency	Infrequent	frequent

Data-driven VVO can effectively address the challenges posed by DERs through statistical, machine learning, reinforcement learning, and hybrid models. Statistical and machine learning models, such as time-series forecasting and probabilistic models, predict renewable power generation. Techniques, such as Monte Carlo simulations, generate a range of possible scenarios. Unsupervised learning techniques can categorize these scenarios into clusters. RLbased models dynamically adapt to variations in renewable power generation, continuously updating their policy networks based on real-time data to make optimization decisions in environments with high penetration of renewable energy sources. Hybrid models combine multiple data-driven approaches to enhance the robustness of VVO. Table 3 shows the summary of techniques for modeling uncertainties of renewable energy sources in VVO.

IV. DATA-DRIVEN DECISION FRAMEWORK FOR VVO

Data-driven VVO uses data-driven approaches, statistics or machine learning, to determine optimal control actions for voltage and reactive power regulations (Fig. 5 and Table 4). The machine learning-based VVO can be supervised, unsupervised, ensemble, and reinforcement learning-based. RL-based VVO has been extensively studied, so it is our main focus in this review.

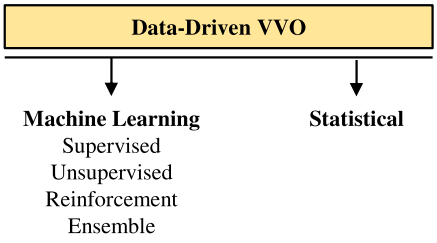


FIGURE 5. The data-driven VVO categories.

A. STATISTICAL MODEL-BASED VVO

Unlike rule-based or pure analytical models, statistical model-based VVO utilizes historical data and various statistical techniques to model, predict, and optimize the behavior of power grids. Statistical techniques applied in VVO include predictive models and uncertainty quantification.

Statistical models, such as linear and nonlinear regression and time-series models, can be used to predict voltages, loads, and reactive power at various points in a MDN. These predictions can serve as "states" in an optimization problem, which can then be manipulated by control variables, such as transformer taps or reactive power injections, to optimize an objective, such as minimizing power losses or maintaining voltage levels within a desired bound. For instance, the bus voltage can be modeled as a linear function of load and reactive power as follows:

$$V_i = \alpha_0 + \alpha_1 P_i + \alpha_2 Q_i + \epsilon \tag{7}$$

where V_i is the voltage at bus i, P_i and Q_i are active and reactive power at bus i, and ϵ is the error of the linear

TABLE 3. The summary of techniques for modeling uncertainties of renewable energy sources in VVO.

Technique	Description	Application	Benefit	Limitation
Statistical	Utilize historical data to forecast and model uncertainties through methods like time-series	Forecasting power outputs of renewable energy sources	Can handle large datasets; provide probabilistic assessments	May not adapt well to real-time changes
Scenario generation/reduction	Model uncertainties by scenario generation/reduction	Modeling variable renewable power generation	Helps in understanding the impact of renewable variability on power grids	Computationally intensive
RL	Dynamically adapt to RES output changes by continuously updating policy networks with real-time data.	Real-time optimization decisions	Adapts in real-time, suitable for environments with high penetration of renewable energy sources	Requires extensive training and fine-tuning. Noise sensitivity, unable to handle different topology
Hybrid	Combine statistical, machine learning, and RL approaches to provide predictive insights and adaptive control	Managing unexpected renewable energy fluctuations	Enhances robustness and reliability of VVO	Noise sensitivity, unable to handle variant topologies

regression model. The model coefficients are represented by α_0, α_1, and α_2. In [36], linear regression is employed to estimate the load, which is then input into the VVO program. Nonlinear regression can also be used in the VVO program, the relationship between variables is inherently non-linear, for example, we have

$$V_i = \alpha_0 + \alpha_1 P_i^2 + \alpha_2 e^{Q_i} + \epsilon \tag{8}$$

In [37], nonlinear regression is used to collect state variables for a distributed VVO in a MDN with multiple virtual power plants. It is based on a well-known distribution optimization algorithm, the alternating direction multiplier method (ADMM). A centralized VVO using nonlinear regression is applied to an inverter-based DG in [38], and variables estimated via nonlinear regression are incorporated into the VVO framework, resulting in a closed-loop VVO.

Data-driven approaches can quantify uncertainties in a MDN through statistics approaches, and statistical models have been widely used as a complementary module for deterministic and stochastic VVO frameworks. Statistical models could model the uncertainty based on probability density functions. For instance, Refs. [39] and [40] use empirical density functions and kernel density functions to model the uncertainty in the form of probability density functions, respectively, using historical data. Statistical models can also be used to model the uncertainty indirectly. For example, in [41], the uncertainty associated with random variables is modeled in a predefined probability density function, such as the Gaussian function. Then, the Bayesian method is used to update the probability functions for the next time intervals. Time-series models as a statistical model, such as autoregressive integrated moving average (ARIMA) [42], can forecast future deviations. Monte Carlo simulation can also be adapted to be data-driven for random sampling by simulating a wide range of scenarios in MDNs. By feeding these simulated scenarios into a VVO algorithm, uncertainties, such as fluctuating demand or variable renewable energy outputs [43], [44], are accounted for.

Statistical model-based data-driven VVO approaches can be implemented easily and are suitable for small-scale MDNs, but they can be sensitive to the data quality, such as noises and anomalies, and may struggle for real-time adaptation and complex topology variations. They generally do not learn or improve over time, which limits their effectiveness in dynamically changing environments.

B. SUPERVISED LEARNING-BASED VVO

The supervised learning-based VVO frameworks use historical data, the data collected from a distribution network, and control settings for optimal voltage and reactive power dispatch to train the model. Once calibrated, the model can predict near-optimal settings for voltage and reactive power regulations in MDNs (Fig. 6). For instance, considering the input vector, $X_t^i = \left[V_t^i, P_t^{i,PV}, tap_t^i, Q_t^{i,cap}\right]$, where V_t^i is the voltage at bus i, $P_t^{i,PV}$ is active power generated by photovoltaics (PVs), tap_t^i is the tap changer status, and $Q_t^{i,cap}$ is reactive power generated by the capacitor at bus i. The training target is represented by $Y_t^i = \left[Q_t^{i,PV}\right]$. The output of the supervised approach is $\tilde{Y}_t^i = \left[\tilde{Q}_t^{i,PV}\right]$, where $\tilde{Q}_t^{i,PV}$ is an approximation to the optimal reactive power dispatch.

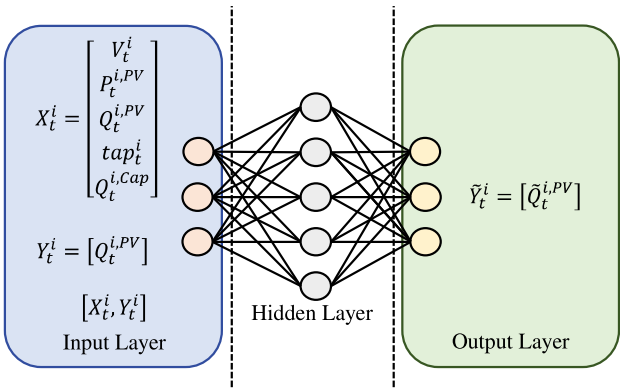


FIGURE 6. The supervised learning-based VVO.

Previous studies on supervised learning-based VVO have used the artificial neural network (ANN), support vector machine (SVM), decision tree, and k-nearest

neighbor (kNN). The switching status of a capacitor bank is determined by ANN in [45]. In [46], ANN controls the voltage magnitude at the point of common coupling (PCC) of PV-connected inverters. Although ANN provides rapid computations, complex relationships between voltage and reactive power in the grid may not be fully captured. SVMs, known for their efficient data use, have been used in [47] for the VVO in distribution networks to coordinate capacitor banks and OLTC settings optimally. In [48], SVM estimates bus voltages and the total power loss. However, SVM may be slow for large datasets, making real-time tasks challenging; SVM also depends on parameter settings, such as kernel functions and regulation parameters. In [49], VVO is formulated as a mixed integer programming (MIP) problem using decision trees. The kNN-based VVO in [50] employs supervised learning to estimate optimal voltage and reactive power for MDNs. By averaging historical voltage and reactive power, optimal settings for a MDN can be forecasted.

These traditional machine learning-based VVOs are computationally efficient, but they may struggle with complex patterns in MDNs. On the other hand, deep learning-based VVO is good at understanding complex patterns by learning directly from raw data, some can even manage time-based data sequences. Deep belief networks (DBNs), convolutional neural networks (CNNs), and recurrent neural networks (RNN) have been used for VVO.

DBNs stack multiple layers of stochastic latent variables and are robust for modeling complex and nonlinear systems. In [51], DBN-based VVO is used to estimate the voltage sensitivity. CNNs are particularly effective in capturing spatial dependencies by using convolutional layers for handling spatial relationships in network configurations. Fig 7 shows a typical CNN-based VVO. CNN has three key parts: convolutional layers, pooling layers, and a fullyconnected layer. In [52], CNN is developed as a local control to regulate the voltage of PV units. CNN is explored as a complementary device in a VVO framework in [53]. Using a temporal CNN, Ref. [54] uses a conventional VVO with capacitor banks and smart inverters. CNN is integrated with control settings of wind turbines (WTs) and PVs in [55] for a VVO in MDNs. The control strategy in [55] is further developed in [56] using the attention mechanism.

RNNs can capture temporal dependencies and are ideal for applications where past states influence the current states. A typical RNN-based VVO is shown in Fig. 8. LSTM networks are a specialized form of RNNs that can remember patterns over long sequences, and can accommodate the temporal dynamics of MDNs. The LSTM and inverter-based VVO framework is presented in [57].

Existing supervised learning-based VVO utilizes labeled historical datasets to calibrate control actions for voltage and reactive power regulations, and can dynamically adapt to fluctuating load demands and integrate DERs seamlessly. However, its precision can be compromised by noises in metering and telemetry data, which may potentially cause the grid instability; its performance may degrade due to significant system topology changes, such as feeder reconfigurations or integration of new substations; it may have overfitting issues, particularly when the model is tuned to a specific grid configuration without considering broader grid dynamics.

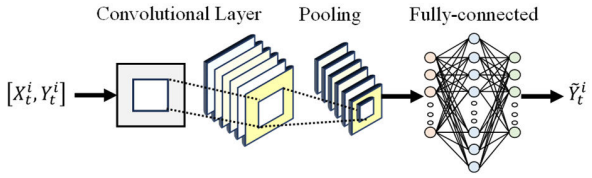


FIGURE 7. A typical CNN-based VVO.

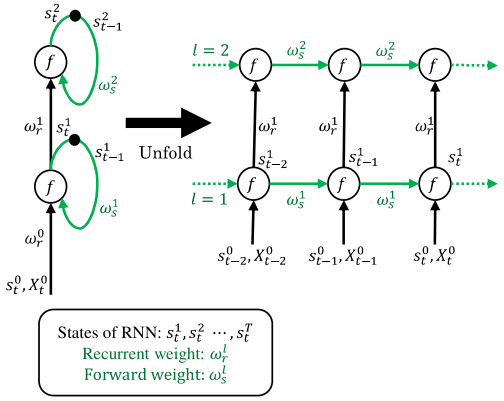


FIGURE 8. A typical RNN-based VVO.

C. UNSUPERVISED LEARNING-BASED VVO

Unlike supervised learning, which uses labeled historical data, the known input-output pairs, to train models (acquiring such labels in power systems can be costly), unsupervised learning does not rely on labeled data. Techniques, such as clustering, can be used to group similar data points or dimensionality reduction to simplify the data's complexity. For instance, kmeans [58], [59] and principal component analysis (PCA) [60] are used to cluster similar generated scenarios using uncertainty quantification approaches. The deep generative adversarial network is used in [61] to solve the uncertainty quantification problem. Unsupervised learning is adept at discovering hidden structures in data, and can be used for the scenario reduction within scenario-based probabilistic VVO frameworks. In the context of VVO, clustering algorithms, such as k-means or hierarchical clustering, can be utilized to segment voltage and VAR data into distinct groups based on their characteristics, and this segmentation aids in identifying common patterns. Unsupervised learning can process vast data from substations and grid sensors efficiently, but without labeled data, voltage and reactive power predictions can be unclear, and the grid stability may not always be ensured. Unsupervised learning can also be sensitive to data noises, may struggle with sudden grid topology changes, and

some algorithms can be computationally demanding. These issues complicate unsupervised learning's real-time applications in VVO, especially for large networks. Unsupervised learning has not been directly applied in VVO.

D. ENSEMBLE LEARNING-BASED VVO

Ensemble learning-based VVO employs multiple models, such as bagging/bootstrap aggregating, boosting, stacking, and voting approaches, with bagging/bootstrap aggregating, such as Random Forest (RF), being the most extensively studied. RF significantly reduces the variance of predictions compared to single models by constructing multiple decision trees and averaging their outputs. This aggregation of diverse model predictions enhances reliability by diluting individual model's biases and errors, leading to more consistent and stable outputs. Furthermore, the ensemble learning improves the prediction accuracy by merging different perspectives from multiple models, which collectively capture a broader spectrum of scenarios and reduce the overall prediction error. In [55], RF-based VVO is compared with the CNN-based VVO framework. Significant topology variations can pose challenges for ensemble learning-based VVO, even with the combined multiple models, recalibration may be required to effectively adapt to grid dynamics and configuration changes.

Statistics, supervised and ensemble learning-based VVO rely a lot on past data, and may not adjust quickly to sudden changes, either in the configuration or behavior, of power systems. However, RL is more flexible and can learn and make decisions based on real-time feedback, making it suitable for VVO.

TABLE 4. Summary of data-driven VVO framework.

VVO	Refs	Attributes (+/–)
Statistics-based	[36]–[44]	+ Easy Implementation + Suitable for small-scale MDNs – Sensitive to noises – High computational complexity – Unable to adapt to topology variation
Supervised learning-based	[45]–[57]	+ Adaptability + Fats Performance – Need large-scale datasets – Noise sensitivity – Unable to handle topology variations
Unsupervised learning-based	[58]–[61]	+ Do not require labeled data + Efficient process – Need large-scale datasets – Noise sensitivity
Ensemble learning-based	[55]	+ Strong learning ability – Computational complexity – Noise sensitive – A large number of hyperparameters

V. RL-BASED DECISION FRAMEWORK FOR VVO

In this section, RL-based decision-making frameworks for VVO, including MDP, MG, CMDP, CMG, and ADMDP, are introduced. In RL-based VVO, voltage and reactive power control is modeled as a MDP that aims to achieve a global Volt/Var control strategy while satisfying the Markov property, which is defined as the probability of being in a certain state at $t+1$ depending only on the previous state and action at t, but not before t. Common elements of a MDP include the state S_t, action A_t, reward R_t, return G_t, agent, and environment, which are defined as follows:

State S_t: A state describes a system's condition at each moment in time. $S_t = \left\{P_{Li}, Q_{Li}, \bar{P}_{Gi}, \bar{Q}_{Gi}, V_i, \sigma, P_{PV}, Q_{PV}\right\}$ is the most frequently utilized conventional state in previous studies. The state space consists of a large number of states, and the number might increase depending on the objective function of VVO. In most cases, this state space is identified by real-time measurements or through the power flow analysis.

Action $\boldsymbol{A_t}$: An action is a move performed by control devices to reach control objectives. Commonly used actions in the literature include: $A_t = \left\{a_t^{CB}, a_t^{OLTC}, a_t^{VR}, a_t^{SVC}, a_t^{STATCOM}, a_t^{SI}\right\}, a_t^{CB} = \{0, 1\}$ (the on-off commitment of CBs), $a_t^{OLTC} = \left\{1, \cdots, N^{OLTC}\right\}$ (N^{OLTC} is the number of tap positions) and $a_t^{VR} = \left\{1, \cdots, N^{VR}\right\}$ (N^{VR} is the number of VR discrete voltage steps). Reactive power injected by SVC and STATCOM is determined by

$$a_t^{SVC/STATCOM} = \frac{Q_t^{SVC/STATCOM}}{\bar{Q}_t^{SVC/STATCOM}} \tag{9}$$

The reactive power injected by SI is based on:

$$a_t^{SI} = \frac{Q_t^{SI}}{\sqrt{\left(S_t^{SI}\right)^2 + \left(P_t^{SI}\right)^2}} \tag{10}$$

$$-1 \leq a_t^{S_t} \leq 1 \tag{11}$$

Reward R_t: A reward is an immediate result at each moment in time that the agent receives after taking an action while interacting with the environment. The reward shows the effectiveness of the agent's decision-making. In RLbased VVO frameworks, the reward depends on the loss function. For instance, for the voltage minimization, we have

$$r_t^{\Delta V} = -\sum_{i=1}^{N^{bus}} ||V_i| - 1| + \varsigma \tag{12}$$

where ς and N^{bus} are the penalty term and the bus number in a MDN, respectively. The reward function for active power losses is defined by

$$r_t^{p_{loss}} = -p_{loss} + \varsigma \tag{13}$$

The reward function for the voltage violation, active power generation, and reactive power consumption are expressed as follows:

$$r_t^{v_{violation}} = -\varsigma \sum_{i=1}^{N^{bus}} \left[1 - min\left(1 - v_{thr} - |1 - V_i|, 0\right)\right]^2 \tag{14}$$

$$r_t^{P_{Gi}} = \varsigma \frac{\sum_{i=1}^{N^{bus}} P_{Gi}}{\sum_{i=1}^{bus} \bar{P}_{Gi}} \tag{15}$$

$$r_t^{Q_{Li}} = \varsigma \frac{\sum_{i=1}^{N^{bus}} Q_{Li}}{\sum_{i=1}^{bus} \bar{Q}_{Li}} \tag{16}$$

Return $\boldsymbol{G_t}$*:* A return is the cumulative reward that the agent obtains at a certain moment in time.

Agent: An agent is a learner and decision-maker who is responsible for interacting with the environment, observing the state, and taking actions to find an optimal control policy that maximizes the expected discounted reward. In most VVO problems, the DSO is the agent.

Environment: The environment includes all aspects of the task that an agent cannot control completely. The environment in the VVO problem of MDNs is the mapping function:

$$f^{Env}\left(P_{Li}, Q_{Li}, P_{Gi}, Q_{Gi}, V_i, \sigma_i\right) \rightarrow \langle V_i, Q_i \rangle \tag{17}$$

Given any state s and action a, the probability of each possible pair of the next state s', and reward r are denoted by the Markov property in (18) [62].

$$p\left(s', r \mid s, a\right) = \Pr\left\{R_{t+1} = r, S_{t+1} = s' \mid S_t, A_t\right\} \tag{18}$$

In a VVO problem, the optimal control strategy aims to mitigate voltage issues, reduce operational costs, and minimize power losses [12]. Fig. 9 shows how the agent and environment interact with each other at each moment in time based on the MDP. First, the agent observes the state S_t from the environment and selects an action A_t. At the next time interval, the agent receives a reward R_{t+1} and observes a new state S_{t+1} as the consequence of its action. Then, the agent takes another action and reaches the next state while getting a new reward. This cycle repeats until the end of the episode [63].

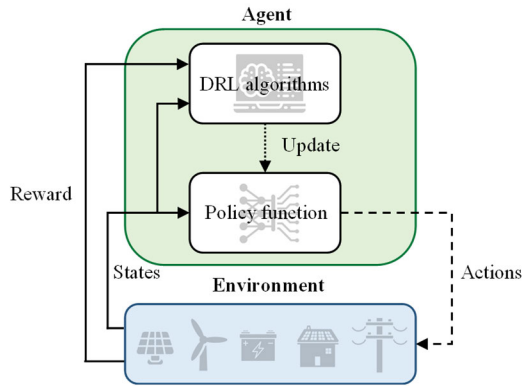


FIGURE 9. The interaction between the agent and environment in the MDP.

In a MDP, the quality of taking actions at each state and time interval is important. A policy is a function that measures this quality and decides what actions to take in a particular state. The input of a policy is the state, and its output is actions. Based on its output actions, policies can be categorized into deterministic and stochastic policies. The deterministic policy π_t maps the state space to the action space, while the stochastic control policy maps the state to the probability of chosen actions in the environment, $\pi_t(a \mid s)$. An optimal policy finds valuable actions in each state that culminate in higher rewards. Therefore, in an RL problem, finding the optimal policy by estimating state- and action-value functions is necessary.

In a MDP, the state-value function $\mathrm{V}_\pi(s)$ is defined as the expected discounted return starting from that state while interacting with the environment following the policy π until the end of the episode in (19). The action-value function $\mathrm{q}_\pi(s, a)$ is the expected discounted return starting from that state and taking an action to interact with the environment following the policy π until the end of the episode in (20). $\mathbb{E}_\pi(.)$ is the expected cumulative reward following the policy π, t is a time interval, and $\gamma \epsilon [0, 1]$ is the discount factor to balance immediate and future rewards ($\gamma = 1$ encourages the agent to consider the long-term consequences of its actions, while $\gamma = 0$ makes the agent short-sighted).

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi\left[G_t \mid S_t = s\right] \\ &= \mathbb{E}_\pi\left[\sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right] \end{aligned} \tag{19}$$

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right] \\ &= \mathbb{E}_\pi\left[\sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right] \end{aligned} \tag{20}$$

The Bellman equations, (19) and (20), are fundamental equations in RL that express the relationship between the value of a state (or the state-action pair) and the values of its successor's state (or the successor state-action pairs). Eq. (21) represents the Bellman equation for the state-value function, which indicates that the value of a state under policy π is the expected immediate reward plus the discounted value of the next state, considering all possible actions and the next state.

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi\left[\sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right] \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma \sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t = s\right] \\ &= \sum\nolimits_a \pi(a \mid s) \sum\nolimits_{s'} \sum\nolimits_r p\left(s', r \mid s, a\right)[r \\ &\quad + \gamma \mathbb{E}_\pi\left[\sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_{t+1} = s'\right]\Big] \\ &= \sum\nolimits_a \pi(a \mid s) \sum\nolimits_{s'} \sum\nolimits_r p\left(s', r \mid s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right] \end{aligned} \tag{21}$$

Similarly, Eq. (22) is the Bellman equation for the action-value function, implying that the value of taking a certain action in a certain state under policy π is the expected immediate reward plus the discounted value of the next state-action pair, considering all possible next states and actions.

$$\begin{aligned} & q_\pi(s) \\ &= \mathbb{E}_\pi\left[\sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right] \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma \sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t = s, A_t = a\right] \\ &= \sum\nolimits_{s'} \sum\nolimits_r p\left(s', r \mid s, a\right)[r \\ &\quad + \gamma \mathbb{E}_\pi\left[\sum\nolimits_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_{t+1} = s', A_{t+1} = a'\right]\Big] \\ &= \sum\nolimits_{s'} \sum\nolimits_r p\left(s', r \mid s, a\right)\left[r + \gamma \sum\nolimits_{a'} \pi\left(a' \mid s'\right) q_\pi\left(s', a'\right)\right] \end{aligned} \tag{22}$$

To solve a MDP problem and find the optimal policy π_* optimal actions are chosen in each state, $v_\pi(s)$ and $q_\pi(s, a)$ are maximized in (23) and (24) for all states and actions.

$$v_{\pi*}(s) = \max v_\pi(s) \tag{23}$$

$$q_{\pi*}(s, a) = \max q_\pi(s, a) \tag{24}$$

A. MARKOV GAME

Fig. 10 shows MG is a multi-agent extension of MDP, where numerous agents collaborate in a common environment to accomplish a goal [64]. Multi-agent RL can be categorized into cooperative, competitive, and a mixture of cooperative and competitive. In the cooperative MG, all agents seek a common goal; in the competitive MG, agents compete with each other toward a goal; in the mixture of cooperative and competitive MG, individuals in the same group coordinate with each other against other groups [65].

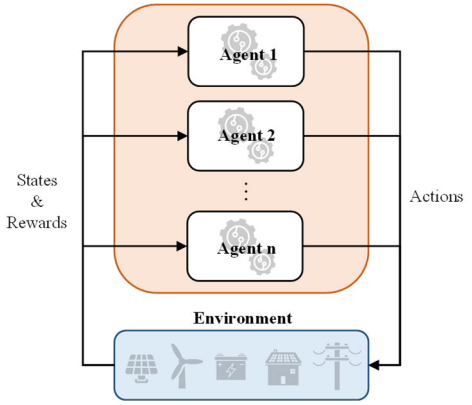


FIGURE 10. The interaction between agents and the environment in a MG.

B. CONSTRAINED MARKOV DECISION PROCESS

In a standard MDP, agents learn their optimal policy by trial and error, but it might jeopardize the safety and reliability of the distribution network since some actions may lead to system divergence and cause the equipment damage. Solving the VVO through CMDP while modeling physical operational constraints, such as power flow limits, voltage deviations, and switching action limits for control devices, improves the safety of a power system. CMDP ensures that every exploration during the learning process is safe by handling two different functions for reward and constraints [66].

C. CONSTRAINED MARKOV GAME

CMG represents an evolving and safer version of MG where multiple agents collaborate under behavioral constraints. In a CMG framework, the safety of executing actions by each agent within a specific state is carefully considered, ensuring that every move is strategically robust and does not jeopardize the integrity of the system or other agents. These behavioral constraints make the decision-making framework more realistic, safe, and applicable [67].

D. ADVERSARIAL MARKOV DECISION PROCESS

AMDP is an expansion of MDP with two learners, known as the protagonist and the adversary (opponent), with opposite goals of adjusting modeling errors. It is suitable for reducing the gap between offline training and online execution while improving the safety and efficiency of control tasks [68]. Table 5 summarizes the RL-based VVO frameworks.

VI. RL-BASED ALGORITHMS FOR VVO

RL-based VVO frameworks are used for sequential decision-making by interacting with the environment to maximize cumulative rewards. Actions leading to rewards are more likely to be repeated, while those causing discomfort or penalties are avoided [62]. Deep reinforcement learning (DRL) combines RL with deep neural networks (DNNs) to address complex tasks with high-dimensional continuous state spaces. DNN performs as a function approximator to tackle the "curse of dimensionality" in such spaces [69]. This is especially relevant in power systems, where state observations are typically continuous values. Fig. 11 shows the main RL and DRL algorithms utilized in VVO, including value-based, policy-based, actor-critic-based, and graph-based algorithms.

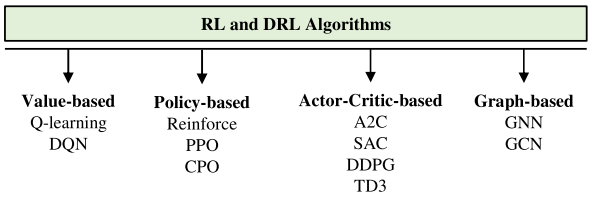


FIGURE 11. The policy-based categories of RL and DRL algorithms in VVO.

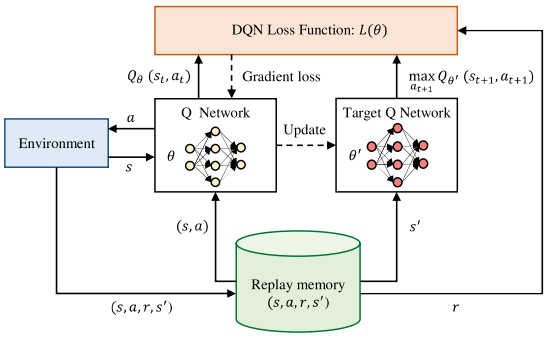


FIGURE 12. The DQN algorithm structure.

A. VALUE-BASED ALGORITHMS

Value-based RL algorithms aim to find an optimal policy by estimating the value function for states or state-action pairs through consistent value updates from observed rewards and agent-environment interactions. The most well-known value-based algorithm, Q-learning, estimates the value of each state-action pair, known as Q-values; when combined with deep learning, it becomes the Deep Q-Network (DQN) [70].

TABLE 5. The summary of RL-based VVO frameworks.

Framework	Advantages	Disadvantages
MDP	Update decisions based on current states and outcomes	Complexity in managing large number of states Require real-time measurements in distribution networks
MG	Allow multi-agent collaboration	Coordination complexity with large number of agents
CMDP	Incorporate physical operational constraints	Limit the exploration space
CMG	Ensure safety and strategic robustness of actions	Increase the complexity due to behavioral constraints
AMDP	Address modeling errors by opposing goals of two learners. Improve safety	Increase the computational complexity due to the dual-learning system

The structure of a DQN is shown in Fig. 12. Q-learning and the DQN face scalability challenges and are impractical for large-scale systems with a large number of control actions and control devices.

B. POLICY-BASED METHODS

Unlike value-based algorithms, policy-based RL methods are designed to directly learn the policy (a function that maps states to actions) and parameterize the policy, so they can handle a large continuous action space. The common approach is to use a neural network (NN) (often named a policy network) to approximate the policy. The input and output of the NN are the state and the probability distribution of taking actions, respectively. This distribution represents the likelihood of taking each action for the given current state. The NN maximizes the expected return and reward for stochastic policy learning, but training policy-based methods is challenging due to the high variance in gradient estimates, which can cause instability during optimization. Common policy-based methods include Reinforce (use Monte Carlo sampling to estimate policy gradients), the proximal policy optimization (PPO) (use a trust region approach for policy updates) [71], and the constrained policy optimization (CPO) algorithm (introduce constraints to ensure safety and stability for policy optimization) [66], [72].

C. ACTOR-CRITIC-BASED ALGORITHMS

Actor-critic (A2C)-based RL algorithms combine features of the value- and policy-based methods. The "actor" determines actions based on the current policy, while the "critic" assesses the quality of these actions by estimating the expected future rewards. By learning from the critic's feedback, the actor can update its policy, thereby, achieving more efficient learning than policy-based methods.

The soft actor-critic (SAC), a maximum entropy framework that encourages exploration, is widely used for VVO (Fig. 13) [73]. Robust policies are achieved by maximizing the policy's entropy, and effectively balancing exploration and exploitation. A SAC algorithm comprises the replay memory, an actor (a policy network), the critics (two Q networks), and loss functions. The replay memory enables the diverse training data by random selection of previous experiences $\left(s, a, r, s'\right)$. The actor, denoted as $\pi\left(a_t \mid s_t\right)$, is a NN that generates probabilistic policy for the given states. The critics are two separate Q networks, denoted as $Q_{\theta_1}\left(s_t, \pi\left(s_t\right)\right)$ and $Q_{\theta_2}\left(s_t, \pi\left(s_t\right)\right)$; they are used by the SAC algorithm to compute Q-values, mitigate the overestimation bias in the value estimation, and determine the most probable action for each state according to the current policy. SAC uses two loss functions for training the actor and critics. For the critics, the loss function, $L\left(\theta_{1,2}\right)$, aims to minimize the discrepancy between the predicted and target Q-values (computed using the policy and Q networks) as follows:

$$L\left(\theta_1\right) = \mathbb{E}_\pi\left[\left(Q_{\theta_1}\left(s_t, \pi\left(s_t\right)\right) - \left(r + \gamma \min_{a_{t+1}}\left(Q_{\theta_1'}\left(s_{t+1}, \pi\left(s_{t+1}\right)\right), Q_{\theta_2'}\left(s_{t+1}, \pi\left(s_{t+1}\right)\right)\right)\right)^2\right] \tag{25}$$

$$L\left(\theta_2\right) = \mathbb{E}_\pi\left[\left(Q_{\theta_2}\left(s_t, \pi\left(s_t\right)\right) - \left(r + \gamma \min_{a_{t+1}}\left(Q_{\theta_1'}\left(s_{t+1}, \pi\left(s_{t+1}\right)\right), Q_{\theta_2'}\left(s_{t+1}, \pi\left(s_{t+1}\right)\right)\right)\right)^2\right] \tag{26}$$

For the actor, the loss function, $L(\pi)$, is the negative of the expected Q-values under the current policy, minus an entropy term, $\log\left(\pi_\varphi\left(a_t \mid s_t\right)\right)$, which encourages exploration in (27).

$$L(\pi) = \mathbb{E}_\pi\left[-\left(Q_{\theta_1}\left(s_t, \pi\left(s_t\right)\right) - \alpha * log\left(\pi_\varphi\left(a_t \mid s_t\right)\right)\right)\right] \tag{27}$$

where α is a temperature parameter that controls the tradeoff between exploitation and exploration. The network parameters are updated via a process, known as "gradient descent", to minimize the respective losses [74].

The deep deterministic policy gradient (DDPG), another popular actor-critic method for VVO, can effectively handle high-dimensional continuous action spaces [75], but is brittle due to sensitivity to hyperparameters [65], i.e., its performance varies significantly depending on the chosen hyperparameters. To overcome this issue, the twin delayed deep deterministic policy gradient (TD3) algorithm is a viable alternative [76]. By using a pair of critics and delaying the policy update, TD3 can address some of DDPG's instability issues.

Overall, actor-critic methods benefit from a combination of value estimation and policy optimization for stable and

efficient learning in RL problems, but may still face issues, such as a high variance and a slow convergence rate.

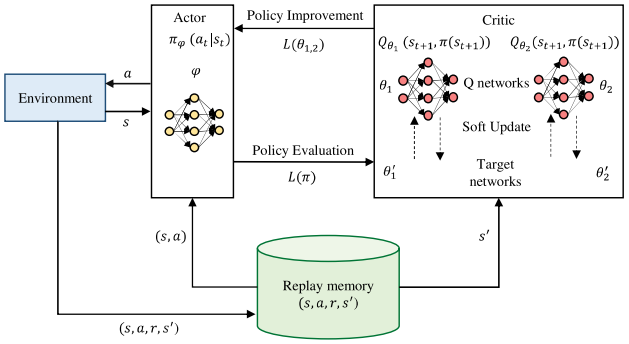


FIGURE 13. The SAC algorithm structure.

D. GRAPH-BASED METHODS

Graph-based RL methods use a graphical representation of the state or action space for complex problems with an inherent network or graphical structure. The environment is represented as a graph, where nodes represent states and edges signify possible actions and transitions. Through a structured way to manage a complex space, the learning efficiency can be significantly improved. A popular graph-based method is the graph neural network (GNN) designed to learn from graph-structured data and capture the relationship among different nodes. GNNs can handle highdimensional data efficiently, capture inter-dependencies among control devices for VVO, and improve the action selection accuracy. Graph convolutional networks (GCNs) are also suitable for VVO, where the state-action space can be naturally depicted as a graph [77]; they can learn to predict future states of a power grid, adjust control policies, detect anomalies, and monitor the health of grid components. However, GNNs and GCNs require more computational resources, especially for large graphs, and require expert knowledge to design an appropriate graph representation of the environment.

RL-based VVO algorithms are summarized in Table 6.

VII. RL-BASED VVO

A. MDP FOR VVO

RL-based VVO frameworks ensure replicable benchmarking and validity of comparative results. In [78], an open-source RL environment, known as PowerGym, for VVO in MDNs is proposed, which provides three types of IEEE standard test systems with CBs, VRs, and ESSs. To validate the proposed environment, a VVO problem is defined by MDP and solved by two RL algorithms, PPO and SAC (SAC converges faster).

MDP-based RL algorithms have been used to improve VVO in distribution networks. The Q-learning algorithm is used in [79] to solve VVO, the set point of SRCDs (CBs and OLTCs) is determined, and the system operational constraints are satisfied. Tap positions of OLTCs for VVO are determined by MDP and batch RL algorithm in [80], and a linearized power flow model is used to estimate the voltage at each node with different tap settings. MDP-based VVO is proposed to schedule reactive power of SIs connected to PVs and ESSs through DDPG, TD3, and SAC algorithms [81], [83]. Operational costs are incorporated into the objective function in [83]. In [84], an entropy-regularized RL-based real-time VVO is used in a wind farm to minimize voltage deviations and power losses by optimizing operations of SIs and SVCs, showing better stability, optimality, and convergence speed than DDPG and TD3. SAC is implemented to optimize reactive power supplied by SIs (connected to PVs and WTs) and SVCs in a distribution network with high voltage variations caused by renewable energy sources [85]; SAC is also used to optimize the schedule of SVCs and PV's SIs [86]. Ref. [87] proposes a mean-field RL (MFRL) algorithm to address scalability issues of standard value-based RL algorithms by adopting the mean-field theory to iteratively discover the agent response to neighboring agents and approximate agent interactions by averaging effects of neighboring agents. In [88], a graph-based proximal policy optimization (GraphPPO) algorithm is developed for VVO, and compared to conventional policy-based methods with dense networks (Dense-PPO); Graph-PPO is more robust than Dense-PPO. Coordinating SRCDs and FRCDs at different timescales through MDP is essential to improve VVO. A two-timescale voltage control method is proposed in [89] to coordinate SRCDs (CBs) and FRCDs (PV's SIs); the on-off status for CBs is determined through MDP-DQN at a slow timescale, while the set point of SIs is determined at a fast timescale using the model-based optimization through a linearized power flow model. In [90], a two timescale VVO is proposed to coordinate SRCDs (OLTC, CBs, and VRs) and FRCDs (PV SIs) in a distribution network. SRCDs are controlled by a model-based approach formulated as a mixed-integer nonlinear programming (MINLP) problem at a slow timescale (hourly); FRCDs are controlled by the DDPG algorithm at a fast timescale (every minute), showing low line losses, voltage deviations, and active power curtailment. In [91], a two-stage voltage regulation framework is proposed to coordinate SRCDs (OLTCs and CBs) and FRCDs (PV's SIs) at different timescale operations. In the day ahead stage, a mixed integer second order cone optimization programming (MISOCP) model is proposed to dispatch SRCDs; in the real time stage, a graph convolutional network-based DDPG (GCN-DDPG) is proposed to regulate reactive power provided by SIs. It outperforms the fully connected network-based and CNN-based DDPG methods in reducing voltage variations. Ref. [92] presents a two-timescale RL-based VVO to coordinate SRCDs (CBs) and FRCDs (ESSs and SIs). The optimal scheduling of CBs with a discrete action space is done by DQN for every three hours; optimal reactive power of SIs, active and reactive power of ESSs in a continuous action space are determined by DDPG for every 30 minutes.

TABLE 6. The Summary of RL-based VVO algorithms.

Algorithm Category	Main Features	Advantages	Disadvantages
Value-based	Estimates the value function for state-action pairs	Simplifies decision-making by evaluating direct outcome	Faces scalability issues in large systems Struggle with a high number of actions and states.
Policy-based	Directly learns the policy that maps states to actions; utilizes neural networks for policy approximation.	Handles large and continuous action spaces efficiently. Capable of learning complex policy structures.	High variance in gradient estimates can cause instability.
Actor-Critic-based	Combines value and policy-based methods; uses an "actor" to choose actions and a "critic" to evaluate them.	Balances exploration and exploitation well	May still experience high variance and slow convergence Complexity in balancing actor and critic learning rates
Graph-based	Utilizes graphical representations of state or action spaces; includes methods l	Effectively manages complex spaces with structured relationships	Requires significant computational resources needs expert knowledge to design graph representations.

TABLE 7. Taxonomy of MDP-based VVO in MDNs.

Refs	Algorithm	Type	Advanced Control Technologies									Objective
			CB	OLTC	VR	DG	ES	SVC	SI	HDT	DR	
[78]	SAC	Actor-critic-based	✓	-	✓	-	✓	-	-	-	-	Voltage deviation, power loss, switching cost
[79]	Q-learning	Value-based	✓	✓	-	-	-	-	-	-	-	Variation of Operation constraints
[80]	Batch RL	Value-based	-	✓	-	-	-	-	-	-	-	Voltage deviation
[81]	DDPG	Actor-critic-based	-	-	-	-	✓	-	✓	-	-	Voltage deviation, power loss
[82]	TD3	Actor-critic-based	-	-	-	-	✓	-	✓	-	-	Voltage deviation, power loss
[83]	SAC	Actor-critic-based	-	-	-	-	✓	-	✓	-	-	Voltage deviation, power loss, operation cost
[84]	entropy-regularized RL	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation, power loss
[85]	SAC	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation, power loss cost
[86]	SAC	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation, power loss cost
[87]	MFRL	Value-based	✓	-	✓	-	-	-	✓	-	-	Voltage deviation, operation cost
[88]	Graph-PPO	Graph-based	✓	-	✓	-	✓	-	-	-	-	Voltage deviation, power loss, switching cost
[89]	DQN	Value-based	✓	-	-	-	-	-	-	-	-	Voltage deviation
[90]	DDPG	Actor-critic-based	-	-	-	-	-	-	✓	-	-	Line loss, voltage violations, active power curtailment cost, the inverter degradation cost
[91]	GCN-DDPG	Graph-based	-	-	-	-	-	-	✓	-	-	Voltage deviation
[92]	DQN & DDPG	Value-based & Actor-critic-based	✓	-	-	-	✓	-	✓	-	-	Power loss, inverter loss, voltage deviation, operation cost
[93]	SBDDPG	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation, PV curtailment
[94]	PISMSAC	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation, power loss

Surrogate models can enhance VVO by simplifying computations. In [93], a surrogate model-based DDPG algorithm (SBDDPG) is proposed to reduce voltage deviations and the PV curtailment of a three-phase unbalanced distribution network, where a supplementary feed forward deep NN is trained in a supervised manner as a surrogate model to learn the relationship between the voltage, active and reactive power at each node. The DDPG algorithm is then used to train VVO from historical data while calculating the reward from the surrogate model. The performance of SBDDPG is similar to conventional DDPG algorithms with accurate system information. In [94], a physics-informed surrogate model-based SAC algorithm (PISMBSAC) is proposed for VVO, which is robust to anomalous measurements without relying on network parameters. A physics-informed global graph attention network (GGAT) and a deep auto-encoder (DAE) network are first used for feature extraction, SAC is then applied for voltage control, and GGAT is applied as a surrogate model to approximate the power flow computation and provide a reward signal. Table 7 shows the summary of MDP-based VVO in the literature.

B. MG FOR VVO

In response to the need for multi-agent RL-based VVO, GridLearn (a flexible framework that extends the opensource CityLearn platform) for grid-level and an OpenAIGym environment for building-level are introduced in [95] to train multi-agent RL algorithms. Multi-agent PPO (MAPPO) is used to address voltage regulations with high- penetration of renewable energy sources along with SIs, ESSs, and DR to enhance grid stability.

MG-based RL algorithms for VVO in MDNs have been studied in the literature. In [64], a decentralized multi-agent MG-based VVO is proposed to optimize large-scale power grids using DDPG. A multi-agent DDPG (MADDPG) algorithm is utilized in [96] and [97]. A MG framework is proposed in [96] to solve VVO through a MADDPG algorithm, which is integrated with an attention model for a system

TABLE 8. Taxonomy on MG-based approaches for VVO in MDNs.

Ref	Algorithm	Type	Advanced Control Technologies									Objective
			CB	OLTC	VR	DG	ES	SVC	SI	HDT	DR	
[95]	MAPPO	Policy-based	-	-	-	-	✓	-	✓	-	✓	Voltage deviation
[64]	MADDPG	Actor-critic-based	-	-	-	✓	-	-	-	-	-	Voltage deviation
[96]	MADDPG	Actor-critic-based	-	-	-	-	-	-	✓	-	-	Voltage deviation
[98]	MADDPG	Actor-critic-based	-	-	-	✓	✓	-	✓	-	✓	Power loss, operating cost
[97]	MADDPG-TSPT	Actor-critic-based	-	-	-	-	-	-	✓	-	-	Voltage deviation
[99]	MSAC	Actor-critic-based	-	-	-	-	-	-	✓	-	-	Voltage deviation, PV curtailment
[100]	MSAC	Actor-critic-based	-	-	-	-	-	-	✓	-	-	Voltage deviation, power loss
[101]	MSAC	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation, power loss
[102]	MATD3	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation
[103]	MOSTC-DP	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Voltage deviation, power loss
[104]	DQN	Value-based	✓	-	✓	-	-	-	✓	-	-	Voltage deviation, power loss
[105]	C-MARL	Actor-critic-based	✓	✓	✓	-	-	-	-	-	-	Operating cost of controllable devices, power loss
[106]	MATRPO	Policy-based	✓	-	✓	-	-	-	✓	-	-	Power loss, voltage deviation
[107]	MADDPG	Actor-critic-based	✓	✓	-	-	-	-	✓	-	-	Voltage deviation, power loss
[109]	SAC & MASAC	Actor-critic-based	✓	✓	-	-	-	-	✓	-	-	Voltage deviation, switching cost
[110]	SAC & MDSAC	Actor-critic-based	✓	-	✓	-	-	-	✓	-	-	Voltage deviation, power loss, switching cost
[108]	DDPG & MSAC	Actor-critic-based	✓	✓	✓	-	-	-	✓	-	-	line loss, voltage violation cost, inverter degradation cost, operation cost

with high penetration of PVs. The attention model addresses the scalability issue by increasing the number of control variables to minimize voltage deviations. In [98], a MADDPG algorithm for real-time VVO is proposed, considering uncertainties of renewable energy sources, load, electricity prices, and control technologies for dispatchable DGs, SIs, ESSs, and the DR program on flexible loads. A cooperative MADDPG with a novel two-stage progressive training (TSPT) strategy is developed in [97] for VVO with a high pentation of PVs, leading to the improved training speed and convergence. All agents are trained separately in the first stage, and coordinately in the second stage.

A MG framework is proposed in [99] to optimize PV's SIs for voltage regulations and the real power curtailment through SAC. In [100] and [101], a multi-agent SAC (MASAC) algorithm-based real-time VVO is proposed to regulate SIs for PVs and WTs. Compared to MADDPG, SAC, and model-based VVO, the proposed method is effective in mitigating voltage violations and reducing power losses. In [102], a MG framework based on an attention mechanism for VVO is developed to optimize the reactive power of SIs and SVC through the MATD3 algorithm, assisting each agent to focus on the most relevant information to its reward. In [103], a one-step two-critic DRL (OSTC-DRL) approach for both single- and multiagents is presented to optimize reactive power of SIs in MDNs using a deterministic policy algorithm (OSTC-DP).

The MG-based VVO in an unbalanced distribution network is developed in [104] to simultaneously minimize voltage deviations and power losses; the ZIP load model is used, and the power flow is solved using the backward-forward sweep method. In [105], a networked multi-agent MDP is proposed to solve the VVO problem through a consensus multi-agent RL (C-MARL) algorithm, which reduces the amount of data required and improves the communication strategy. In [106], a non-cooperative MG framework is formulated to optimize CBs, VRs, and SIs (for PVs and WTs) using a decentralized multi-agent trust region policy optimization (MATRPO) algorithm while considering uncertainties of load and RERs power generation.

Coordinating SRCDs and FRCDs at different timescales through MG is essential [107], [108]. In [107], a multi-timescale MG-based VVO framework is proposed to coordinate SRCDs (CBs and OLTCs) and FRCDs (SIs) through MADDPG. In [109], a model-free two-timescale voltage control method is provided to control SRCDs (CBs and OLTCs) and FRCDs (SIs). The SRCDs are formulated through a single agent MDP and solved by SAC with an hourly time interval; SIs are coordinated as MG in a smaller time interval and solved by MASAC to address fast voltage fluctuations. In [110], a bi-level off-policy RL method is proposed to jointly coordinate SRCDs (CBs and VRs) and FRCDs (SIs) using the multi-timescale off-policy correction (MTOPC) technique. For the fast timescale, the optimal schedule of SIs is obtained by a single agent trained by SAC; for the slow timescale, the multi-discrete SAC (MDSAC) algorithm is used to optimize the set point of CBs and VRs. In [108], a two-timescale VVO is proposed to jointly optimize SRCDs (CBs, OLTCs, and VRs) and FRCDs (SIs). SIs in the fast timescale are modeled as MDP and optimized by DDPG; traditional control devices are modeled as MG and optimized by the MASAC algorithm in a slow timescale. The two policies are interconnected by a communication protocol and are learned concurrently.

However, MDP and MG either consider constraints as a penalty in the objective function or avoid considering constraints, so online execution of the learned strategy may not be practical in real life due to constraint violations. CMDP and CMG are more effective by incorporating constraints into the learning process. Table 8 shows the summary of MG-based VVO in the literature.

TABLE 9. Taxonomy on CMDP-based approaches for VVO in MDNs.

Ref	Algorithm	Type	Advanced Control Technologies									Objective
			CB	OLTC	VR	DG	ES	SVC	SI	HDT	DR	
[111]	CSAC	Actor-critic-based	✓	✓	✓	-	-	-	-	-	-	Operating cost of controllable devices, power loss
[112]	CPO	Policy-based	✓	✓	✓	-	-	-	-	-	-	Operating cost of controllable devices, power loss
[113]	SAAC	Actor-critic-based	✓	✓	✓	-	-	-	-	-	-	Operating cost of controllable devices, power loss, voltage deviation
[114]	DDPG	Actor-critic-based	-	-	-	-	-	-	-	✓	-	Power loss, voltage regulation
[115]	CPO	Policy-based	✓	✓	✓	✓	✓	-	-	-	-	Operation cost, adaptive voltage regulation
[116]	SSAC	Actor-critic-based	-	-	-	-	✓	-	✓	-	-	Voltage deviation, power loss

C. CMDP FOR VVO

CMDP-based VVO demonstrates its remarkable ability to navigate operational constraints while optimizing multiple control devices. In [111], VVO is modeled as CMDP to optimize SRCDs (CBs, OLTCs, and VRs) by minimizing operation costs while satisfying the bus voltage constraints. The developed model is solved by CSAC, and compared with DQN, SAC, CPO, and optimization-based approaches, showing better sample efficiency, scalability, and constraint satisfaction. In [112], VVO is modeled as CMDP by implementing two policy-gradient methods, the trust region policy optimization (TRPO) and CPO, to optimize CBs, VRs, and OLTC; CPO outperforms TRPO with negligible voltage violations. In [113], a safety layer augmented actorcritic (SAAC) algorithm for VVO is proposed to address two common RL-based VVO challenges: sample efficiency and safety. To improve the sample efficiency, a modelaugmented partwise derivative technique is incorporated to train the RL algorithm; to improve safety, the actor is equipped with a constraint satisfaction layer based on iterative quadratic programming. A unique mutual information regularizer is then proposed to boost the performance of the constraint satisfaction layer. In [114], VVO is formulated as a CMDP to optimize reactive power generated by HDTs through DDPG, where a safe exploration approach is proposed as a safety layer on top of the policy gradient actor to consider operational constraints. It can immediately respond to voltage fluctuations and significantly reduce power losses.

Coordination of control devices with continuous and discrete action spaces in VVO is important to ensure a more efficient and robust voltage regulation and energy cost minimization. A comprehensive approach is needed with a suitable RL algorithm to coordinate SRCDs and FRCDs with multi-timescale operation. In [115], a VVO problem is formulated as CMDP to adapt voltage regulation and minimize energy costs by optimizing various control devices, including CBs, OLTC, VRs, DGs, and ESSs. To satisfy constraints at each time step during the learning process, CPO is used; to handle discrete and continuous action spaces, a hybrid action space through a stochastic control policy is defined. The designed policy does not have a scalability issue since actions are generated by sampling from a joint distribution of mixed random variables. In [116], a three-stage multi-timescale VVO framework is proposed to coordinate SIs and ESSs. In the first stage, the schedule of ESSs with a 30 -minute resolution is optimized for peak shaving; in the second stage, a safe SAC (SSAC) algorithm is developed to coordinate ESSs and PV SIs to minimize voltage deviations and power losses within a 1-minute resolution; in the third stage, a proportional-integral (PI) controller supplemented by real power compensation is integrated into SIs (for PVs and ESSs) to quickly address voltage deviations in a 0.1-second resolution. Table 9 shows the summary of CMDP-based VVO in the literature.

D. CMG FOR VVO

CMG-based VVO facilitates safe multi-agent RL in MDNs through effective collaboration and optimization among multiple agents while ensuring that actions taken do not jeopardize the overall stability and functionality of the system. In [77], a safe multi-agent primal-dual graph reinforcement learning (MAPDGRL) approach is proposed to optimize PV's SIs based on a decentralized VVO in a zoned distribution network. GCN can capture graph-based characteristics from distribution networks. This process highlights connections between VVO and the grid topology while effectively eliminating noises and imputing missing data. In [117], an online decentralized multi-agent RL framework is formulated as CMG to solve VVO and optimize the reactive power of control devices through MACSAC by using local measurement data without relying on real-time peer-to-peer (P2P) communication, which is hard to obtain. The proposed method outperforms modelbased methods, MADDPG and CSAC algorithms, for online applications. In [118], a novel CMG framework is proposed by integrating a physics-shield multi-agent twin delayed deep deterministic policy gradient (Physics-shield MATD3) algorithm to provide safe scheduling of ESSs, SIs, and SVCs. The physics-based shielding mechanism helps the agent to replace dangerous actions with safe actions while maintaining the system stability.

Solving CMDPs and CMGs in VVO is computationally complex as constraints are incorporated into the decisionmaking process, so the algorithm must handle constraints while searching for an optimal policy; it also leads to a more fragmented and discontinuous action space, so finding an optimal policy is challenging, and the learning process is more

TABLE 10. Taxonomy on CMG and AMDP-based VVO in MDNs.

Ref	Framework	Agent	Algorithm	Type	Advanced Control Technologies									Objective
					CB	OLTC	VR	DG	ES	SVC	SI	HDT	DR	
[77]	CMG	Multi	MAPDGRL	Graph-based	-	-	-	-	-	-	✓	-	-	Power loss, voltage deviation
[117]	CMG	Multi	MACSAC	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Power loss, voltage deviation
[118]	CMG	Multi	Physics-shielded MATD3	Actor-critic-based	-	-	-	-	✓	✓	✓	-	-	Voltage regulation penalty from the shield, the reactive power generation of SVCs
[68]	AMDP	Single	JASAC	Actor-critic-based	-	-	-	-	-	✓	✓	-	-	Power loss, voltage deviation

time-consuming and prone to convergence issues. Therefore, CMDPs and CMGs may require more extensive computational resources and longer computation time than traditional MDPs, which may limit their real-time applications.

E. AMDP FOR VVO

AMDP-based VVO offers enhanced safety and efficiency by incorporating adversarial learning techniques in decision-making scenarios, suitable for reducing the gap between offline training and online execution. In [68], a two-stage DRL method with offline and online stages is proposed to minimize voltage deviations and power losses through SVCs and SIs. The offline stage for VVO is modeled as AMDP and solved by the Jointly Adversarial Soft Actor-Critic (JASAC) algorithm. Table X shows the summary of CMG and AMDP-based VVO in the literature.

F. OVERVIEW ON RL-BASED VVO FRAMEWORK

RL-based VVO frameworks offer significant advancements over traditional model-based optimization techniques by more effectively addressing dynamics and stochastic events in modern distribution networks. Traditional methods often require accurate and extensive system modeling, but RLbased VVO operates through a trial-and-error learning process that continuously adapts to changes in the network, which enhances the flexibility and responsiveness to real- time grid operating conditions. However, RL-based strategies also have challenges as they typically cannot handle the network topology changes, and such changes frequently occur during the operation of MDNs. DRL-based VVO frameworks perform relying on the observed states. In most previous work, states are estimated based on the modelbased power flow, which may not consider the network topology changes. Another challenge is the data required to be shared among all entities in a distribution network may have different owners with conflict interests, and current DRL-based methods cannot preserve data privacy. DRL also requires substantial computational resources for training and a well-designed reward system to ensure convergence on effective policies. Furthermore, DRL-based frameworks are not robust to noises, particularly non-Gaussian noises (the majority of noises in power systems do not follow Gaussian distributions [119]), and communication system malfunctions.

VIII. FUTURE RESEARCH DIRECTIONS

The data-driven VVO is currently at the research stage and needs significant effort before it can be implemented in real life. The future research directions in this area are recommended below.

Modeling Complexity: Data-driven VVO models are complex to address nonlinearities and uncertainties of a system, and accommodate operational constraints and numerous variables. Integrating RL with other AI techniques, such as evolutionary algorithms and swarm intelligence, may offer improvements.

Grid Topology Changes: A distribution network's topology changes frequently due to planned maintenance, equipment failure, renewable energy sources, and EV charging stations integration. Data-driven VVO models must be adaptive to such topology changes. Incorporating the grid topology information into the state representation of a DRL model or rapidly retraining DRL models after the topology changes is recommended.

Scalability: Data-driven methods tend to have scalability issues for large power systems due to the high computational burden and long training time. State-of-the-art data-driven methods, such as distributional reinforcement learning, implicit quantile networks and dueling architectures, can be used to better manage large power grids.

Data Privacy: Data-driven VVO models need operational data, which may raise confidentiality and security concerns for user data. Developing techniques, such as federated learning, for decentralized learning can ensure data privacy.

Real-Time Implementation: It is important to bridge the gap between offline training and real-time implementation of DRL-based VVO models to ensure models adapt to real-time changes in power grids. The approximate dynamic programming (ADP) and adversarial learning (AD) methods can be explored.

Transfer Learning: Transfer learning allows data-driven VVO models to learn knowledge from one scenario to another, and reduces the dependence on a vast amount of data, so is suitable for scenarios with limited data.

Robustness: Power grids are susceptible to sensor and communication failures, and other sources of uncertainties. Developing data-driven models that work effectively with incomplete or imperfect data is a promising research direction.

Computational Efficiency: Data-driven models have high computational demand, which may prevent effective realtime or near-real-time decision-making. Future research should focus on improving the computational efficiency through the model simplification, process optimization, distributed computing resource leverage, and surrogate modeling.

Noise Signals Handling: Noises in the data can adversely affect a VVO model, which may complicate the learning process and cause suboptimal decisions. It is recommended to integrate noise filtering/smoothing techniques into the

preprocessing stage of a data-driven model, develop robust data-driven algorithms inherently resistant to noises, incorporate uncertainties into the learning process, design state representations that are less sensitive to noises, and develop novel learning algorithms to mitigate the noise impact.

IX. CONCLUSION

VVO is an essential technique for regulating voltage and minimizing power losses in MDNs, especially with a high penetration of DERs. VVO can be broadly categorized into model-based and data-driven methods. Based on the comprehensive literature review conducted in this paper, approximately 69.6% of the reviewed papers focusing on model-based VVO, and 30.4% of the reviewed papers focusing on data-driven VVO. In this paper, we have conducted an in-depth survey of the data-driven VVO using statistics and machine learning techniques, such as supervised, unsupervised, ensemble, and reinforcement learning. The special focus is on model-free RL methods, including the MDP, MG, CMDP, CMG, and AMDP. The coordination of SRCDs and FRCDs across different time scales for VVO applications is also summarized. Future research directions in this important area are recommended to advance the efficiency and effectiveness of VVO strategies in MDNs.

REFERENCES

[1] S. Allahmoradi, M. P. Moghaddam, S. Bahramara, and P. Sheikhahmadi, "Flexibility-constrained operation scheduling of active distribution networks," *Int. J. Electr. Power Energy Syst.*, vol. 131, Oct. 2021, Art. no. 107061, doi: 10.1016/j.ijepes.2021.107061.

[2] *International Renewable Energy Agency (IRENA)*, World Energy Transitions Outlook, Abu Dhabi, United Arab Emirates, 2022.

[3] G. Wang, V. Kekatos, A. J. Conejo, and G. B. Giannakis, "Ergodic energy management leveraging resource variability in distribution grids," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4765–4775, Nov. 2016, doi: 10.1109/TPWRS.2016.2524679.

[4] M. Tahir, M. E. Nassar, R. El-Shatshat, and M. M. A. Salama, "A review of Volt/Var control techniques in passive and active power distribution networks," in *Proc. IEEE Smart Energy Grid Eng. (SEGE)*, Aug. 2016, pp. 57–63, doi: 10.1109/SEGE.2016.7589500.

[5] A. Mahendru and P. Deshpande, "A review of Volt Var optimization techniques," *J. Electr. Electron. Syst.*, vol. 7, no. 3, pp. 1–14, 2018.

[6] V. A. Evangelopoulos, P. S. Georgilakis, and N. D. Hatziargyriou, "Optimal operation of smart distribution networks: A review of models, methods and future research," *Electric Power Syst. Res.*, vol. 140, pp. 95–106, Nov. 2016, doi: 10.1016/j.epsr.2016.06.035.

[7] H. Sun, Q. Guo, J. Qi, V. Ajjarapu, R. Bravo, J. Chow, Z. Li, R. Moghe, E. Nasr-Azadani, U. Tamrakar, G. N. Taranto, R. Tonkoski, G. Valverde, Q. Wu, and G. Yang, "Review of challenges and research opportunities for voltage control in smart grids," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2790–2801, Jul. 2019, doi: 10.1109/TPWRS.2019.2897948.

[8] S. Satsangi and G. B. Kumbhar, "Review on Volt/Var optimization and control in electric distribution system," in *Proc. IEEE 1st Int. Conf. Power Electron., Intell. Control Energy Syst. (ICPEICES)*, Jul. 2016, pp. 1–6, doi: 10.1109/ICPEICES.2016.7853324.

[9] K. E. Antoniadou-Plytaria, I. N. Kouveliotis-Lysikatos, P. S. Georgilakis, and N. D. Hatziargyriou, "Distributed and decentralized voltage control of smart distribution networks: Models, methods, and future research," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2999–3008, Nov. 2017, doi: 10.1109/TSG.2017.2679238.

[10] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, Jan. 2020, doi: 10.1109/TPWRS.2019.2941134.

[11] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE J. Power Energy Syst.*, vol. 5, no. 1, pp. 1–10, Mar. 2019, doi: 10.17775/CSEEJPES.2018.00840.

[12] R. Diao, Z. Wang, D. Shi, Q. Chang, J. Duan, and X. Zhang, "Autonomous voltage control for grid operation using deep reinforcement learning," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2019, pp. 1–5, doi: 10.1109/PESGM40551.2019.8973924.

[13] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive power system emergency control using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1171–1182, Mar. 2020, doi: 10.1109/TSG.2019.2933191.

[14] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Reinforcement learning and its applications in modern power and energy systems: A review," *J. Modern Power Syst. Clean Energy*, vol. 8, no. 6, pp. 1029–1042, Nov. 2020, doi: 10.35833/MPCE.2020.000552.

[15] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, Mar. 2020, doi: 10.17775/CSEEJPES.2019.00920.

[16] M. Glavic, "Deep reinforcement learning for electric power system control and related problems: A short review and perspectives," *Annu. Rev. Control*, vol. 48, pp. 22–35, May 2019, doi: 10.1016/j.arcontrol.2019.09.008.

[17] D. Hai, T. Zhu, S. Duan, W. Huang, and W. Li, "Deep reinforcement learning for Volt/Var control in distribution systems: A review," in *Proc. 5th Int. Conf. Energy, Electr. Power Eng. (CEEPE)*, Apr. 2022, pp. 596–601, doi: 10.1109/CEEPE55110.2022.9783357.

[18] Q. Li, Y. Zhang, T. Ji, X. Lin, and Z. Cai, "Volt/Var control for power grids with connections of large-scale wind farms: A review," *IEEE Access*, vol. 6, pp. 26675–26692, 2018, doi: 10.1109/ACCESS.2018.2832175.

[19] H. Mataifa, S. Krishnamurthy, and C. Kriger, "Volt/Var optimization: A survey of classical and heuristic optimization methods," *IEEE Access*, vol. 10, pp. 13379–13399, 2022, doi: 10.1109/ACCESS.2022.3146366.

[20] K. Gholami, M. R. Islam, M. M. Rahman, A. Azizivahed, and A. Fekih, "State-of-the-art technologies for Volt-Var control to support the penetration of renewable energy into the smart distribution grids," *Energy Rep.*, vol. 8, pp. 8630–8651, Nov. 2022, doi: 10.1016/j.egyr.2022.06.080.

[21] *Electrical Power Systems and Equipment-Voltage Ratings*, ANSI Standard C 84.1, 1995.

[22] *Preferred Voltage Levels for AC Systems 0 To 50 000 V*, Standard CAN3-C235-83, 1983.

[23] T. Taylor. (2011). *Introduction to Model-based Volt/VAR Optimization*. Accessed: Feb. 27, 2024. [Online]. Available: https://www.power-grid.com/smart-grid/introduction-to-model-based-Volt-Var-optimization/

[24] H. V. Padullaparti, Q. Nguyen, and S. Santoso, "Advances in Volt-Var control approaches in utility distribution systems," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2016, pp. 1–5, doi: 10.1109/PESGM.2016.7741366.

[25] (2023). *IEEE 1547 Standard for Interconnecting Distributed Resources With Electric Power Systems*. Accessed: May 18, 2023. [Online]. Available: https://grouper.ieee.org/groups/

[26] X. Sun, J. Qiu, and J. Zhao, "Optimal local Volt/Var control for photovoltaic inverters in active distribution networks," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5756–5766, Nov. 2021, doi: 10.1109/TPWRS.2021.3080039.

[27] A. E. Saldaña-González, A. Sumper, M. Aragüés-Peñalba, and M. Smolnikar, "Advanced distribution measurement technologies and data applications for smart grids: A review," *Energies*, vol. 13, no. 14, p. 3730, Jul. 2020, doi: 10.3390/en13143730.

[28] D. D. Giustina and S. Rinaldi, "Hybrid communication network for the smart grid: Validation of a field test experience," *IEEE Trans. Power Del.*, vol. 30, no. 6, pp. 2492–2500, Dec. 2015, doi: 10.1109/TPWRD.2015.2393836.

[29] B. Uluski, "Volt/var control and optimization concepts and issues," *Electr. Power Res. Inst.*, vol. 1, p. 58, Jun. 2011.

[30] Q. Zhang, Y. Guo, Z. Wang, and F. Bu, "Distributed optimal conservation voltage reduction in integrated primary-secondary distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 3889–3900, Sep. 2021, doi: 10.1109/TSG.2021.3088010.

[31] H. J. Liu, W. Shi, and H. Zhu, "Distributed voltage control in distribution networks: Online and robust implementations," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6106–6117, Nov. 2018, doi: 10.1109/TSG.2017.2703642.

[32] Y. Chai, L. Guo, C. Wang, Z. Zhao, X. Du, and J. Pan, "Network partition and voltage coordination control for distribution networks with high penetration of distributed PV units," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3396–3407, May 2018, doi: 10.1109/TPWRS.2018.2813400.

[33] M. Yilmaz and R. ElShatshat, "Zone-oriented 2-stage distributed voltage control algorithm for active distribution networks," *Electric Power Syst. Res.*, vol. 217, Apr. 2023, Art. no. 109127, doi: 10.1016/j.epsr.2023.109127.

[34] H. Sun, Q. Guo, B. Zhang, W. Wu, and B. Wang, "An adaptive zone-division-based automatic voltage control system with applications in China," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1816–1828, May 2013, doi: 10.1109/TPWRS.2012.2228013.

[35] S. Rahimi, M. Marinelli, and F. Silvestro, "Evaluation of requirements for Volt/Var control and optimization function in distribution management systems," in *Proc. IEEE Int. Energy Conf. Exhibition*, Sep. 2012, pp. 331–336, doi: 10.1109/ENERGYCON.2012.6347777.

[36] Z. S. Hossein, A. Khodaei, W. Fan, M. S. Hossan, H. Zheng, S. A. Fard, A. Paaso, and S. Bahramirad, "Conservation voltage reduction and Volt-Var optimization: Measurement and verification benchmarking," *IEEE Access*, vol. 8, pp. 50755–50770, 2020, doi: 10.1109/ACCESS.2020.2979242.

[37] S. Li, W. Wu, and Y. Lin, "Robust data-driven and fully distributed Volt/Var control for active distribution networks with multiple virtual power plants," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 2627–2638, Jul. 2022, doi: 10.1109/TSG.2022.3166274.

[38] T. Xu, W. Wu, Y. Hong, J. Yu, and F. Zhang, "Data-driven inverter-based Volt/Var control for partially observable distribution networks," *CSEE J. Power Energy Syst.*, vol. 9, no. 2, pp. 548–560, Mar. 2023, doi: 10.17775/CSEEJPES.2020.05920.

[39] Y. Liu, M. Wang, X. Zhang, J. Duan, H. Gao, and J. Liu, "Kriging surrogate model enabled heuristic algorithm for coordinated Volt/Var management in active distribution networks," *Electric Power Syst. Res.*, vol. 210, Sep. 2022, Art. no. 108089, doi: 10.1016/j.epsr.2022.108089.

[40] D. Lee, C. Han, and G. Jang, "Stochastic analysis-based Volt–Var curve of smart inverters for combined voltage regulation in distribution networks," *Energies*, vol. 14, no. 10, p. 2785, May 2021, doi: 10.3390/en14102785.

[41] S. Talkington, S. Grijalva, M. J. Reno, and J. A. Azzolini, "Solar PV inverter reactive power disaggregation and control setting estimation," *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4773–4784, Nov. 2022, doi: 10.1109/TPWRS.2022.3144676.

[42] V. B. Pamshetti, S. Singh, A. K. Thakur, and S. P. Singh, "Multi-stage coordination Volt/Var control with CVR in active distribution network in presence of inverter-based DG units and soft open points," *IEEE Trans. Ind. Appl.*, vol. 57, no. 3, pp. 2035–2047, May 2021, doi: 10.1109/TIA.2021.3063667.

[43] A. Majumdar, Y. P. Agalgaonkar, B. C. Pal, and R. Gottschalg, "Centralized Volt–Var optimization strategy considering malicious attack on distributed energy resources control," *IEEE Trans. Sustain. Energy*, vol. 9, no. 1, pp. 148–156, Jan. 2018, doi: 10.1109/TSTE.2017.2706965.

[44] T. S. Vitor and J. C. M. Vieira, "Operation planning and decision-making approaches for Volt/Var multi-objective optimization in power distribution systems," *Electric Power Syst. Res.*, vol. 191, Feb. 2021, Art. no. 106874, doi: 10.1016/j.epsr.2020.106874.

[45] B. Das and P. K. Verma, "Artificial neural network-based optimal capacitor switching in a distribution system," *Electric Power Syst. Res.*, vol. 60, no. 2, pp. 55–62, Dec. 2001, doi: 10.1016/s0378-7796(01)00149-3.

[46] S. Li, Y. Sun, M. Ramezani, and Y. Xiao, "Artificial neural networks for Volt/Var control of DER inverters at the grid edge," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5564–5573, Sep. 2019, doi: 10.1109/TSG.2018.2887080.

[47] D. Salles, A. C. Pinto, and W. Freitas, "Integrated Volt/Var control in modern distribution power systems based on support vector machines," *Int. Trans. Electr. Energy Syst.*, vol. 26, no. 10, pp. 2216–2229, Oct. 2016, doi: 10.1002/etep.2200.

[48] E. Pourjafari and M. Reformat, "A support vector regression based model predictive control for Volt-Var optimization of distribution systems," *IEEE Access*, vol. 7, pp. 93352–93363, 2019, doi: 10.1109/ACCESS.2019.2928173.

[49] Z. Yan, X. Li, H. Zhou, Z. Xie, Y. Wang, and L. Hong, "Learning strategies for Volt-Var optimization with PV inverters and battery energy storage systems in distribution network," in *Proc. 6th Int. Conf. Energy, Electr. Power Eng. (CEEPE)*, May 2023, pp. 65–70, doi: 10.1109/CEEPE58418.2023.10165823.

[50] P. Bagheri and W. Xu, "Assessing benefits of Volt-Var control schemes using AMI data analytics," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1295–1304, May 2017, doi: 10.1109/TSG.2016.2603421.

[51] N. Shi, Graduate, R. Cheng, L. Liu, Z. Wang, Q. Zhang, and M. J. Reno, "Data-driven affinely adjustable robust Volt/Var control," *IEEE Trans. Smart Grid*, vol. 15, no. 1, pp. 247–259, Sep. 2023, doi: 10.1109/TSG.2023.3270112.

[52] X. Sun, J. Qiu, Y. Tao, Y. Ma, and J. Zhao, "A multi-mode data-driven Volt/Var control strategy with conservation voltage reduction in active distribution networks," *IEEE Trans. Sustain. Energy*, vol. 13, no. 2, pp. 1073–1085, Apr. 2022, doi: 10.1109/TSTE.2022.3149267.

[53] X. Sun, J. Qiu, and J. Zhao, "Real-time Volt/Var control in active distribution networks with data-driven partition method," *IEEE Trans. Power Syst.*, vol. 36, no. 3, pp. 2448–2461, May 2021, doi: 10.1109/TPWRS.2020.3037294.

[54] L. Miao, Y. Peng, Z. Li, W. Xi, and T. Cai, "Data-driven Volt/Var control based on constrained temporal convolutional networks with a corrective mechanism," *Electric Power Syst. Res.*, vol. 224, Nov. 2023, Art. no. 109738, doi: 10.1016/j.epsr.2023.109738.

[55] Y. Zhao, G. Zhang, W. Hu, Q. Huang, Z. Chen, and F. Blaabjerg, "Meta-learning based voltage control strategy for emergency faults of active distribution networks," *Appl. Energy*, vol. 349, Nov. 2023, Art. no. 121399, doi: 10.1016/j.apenergy.2023.121399.

[56] Y. Zhao, G. Zhang, W. Hu, Q. Huang, Z. Chen, and F. Blaabjerg, "Meta-Learning based voltage control for renewable energy integrated active distribution network against topology change," *IEEE Trans. Power Syst.*, vol. 38, no. 6, pp. 5937–5940, Nov. 2023, doi: 10.1109/TPWRS.2023.3309536.

[57] B. Li and Q. Xu, "A machine learning-assisted distributed optimization method for inverter-based Volt-Var control in active distribution networks," *IEEE Trans. Power Syst.*, vol. 39, no. 2, pp. 2668–2681, 2023, doi: 10.1109/TPWRS.2023.3279303.

[58] S. Singh, V. B. Pamshetti, and S. P. Singh, "Time horizon-based model predictive Volt/Var optimization for smart grid enabled CVR in the presence of electric vehicle charging loads," *IEEE Trans. Ind. Appl.*, vol. 55, no. 6, pp. 5502–5513, Nov. 2019, doi: 10.1109/TIA.2019.2928490.

[59] S. Singh, S. Veda, S. P. Singh, R. Jain, and M. Baggu, "Event-driven predictive approach for real-time Volt/Var control with CVR in solar PV rich active distribution network," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 3849–3864, Sep. 2021, doi: 10.1109/TPWRS.2021.3057656.

[60] Q. Nguyen, X. Ke, N. Samaan, J. Holzer, M. Elizondo, H. Zhou, Z. Hou, R. Huang, M. Vallem, B. Vyakaranam, M. Ghosal, and Y. V. Makarov, "Transmission-distribution long-term Volt-Var planning considering reactive power support capability of distributed PV," *Int. J. Electr. Power Energy Syst.*, vol. 138, Jun. 2022, Art. no. 107955, doi: 10.1016/j.ijepes.2022.107955.

[61] X. Xu, M. Wang, Z. Xu, and Y. He, "Generative adversarial network assisted stochastic photovoltaic system planning considering coordinated multi-timescale Volt-Var optimization in distribution grids," *Int. J. Electr. Power Energy Syst.*, vol. 153, Nov. 2023, Art. no. 109307, doi: 10.1016/j.ijepes.2023.109307.

[62] S. J. D. Prince, *Understanding Deep Learning*. Cambridge, MA, USA: MIT Press, 2023.

[63] G. Ruan, H. Zhong, G. Zhang, Y. He, X. Wang, and T. Pu, "Review of learning-assisted power system optimization," *CSEE J. Power Energy Syst.*, vol. 7, no. 2, pp. 221–231, Mar. 2021, doi: 10.17775/CSEEJPES.2020.03070.

[64] S. Wang, J. Duan, D. Shi, C. Xu, H. Li, R. Diao, and Z. Wang, "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4644–4654, Nov. 2020, doi: 10.1109/TPWRS.2020.2990179.

[65] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control*, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Eds. Cham, Switzerland: Springer, 2021, pp. 1–10, doi: 10.1007/978-3-030-60990-0_12.

[66] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. 34th Int. Conf. Mach. Learn. ICML*, May 2017, pp. 30–47.

[67] Z. Chen, S. Ma, and Y. Zhou, "Finding correlated equilibrium of constrained Markov game: A primal-dual approach," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–14.

[68] H. Liu and W. Wu, "Two-stage deep reinforcement learning for inverter-based Volt-Var control in active distribution networks," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2037–2047, May 2021, doi: 10.1109/TSG.2020.3041620.

[69] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage regulation in distribution grids using deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids*, Oct. 2019, pp. 1–6, doi: 10.1109/SMARTGRIDCOMM.2019.8909764.

[70] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," 2019, *arxiv:1901.00137*.

[71] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arxiv:1707.06347*.

[72] C. Ying, X. Zhou, H. Su, D. Yan, N. Chen, and J. Zhu, "Towards safe reinforcement learning via constraining conditional value-at-risk," Jun. 2022, *arXiv:2206.04436*, doi: 10.48550/arXiv.2206.04436.

[73] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," 2018, *arxiv:1812.05905*.

[74] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-Policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn. ICML*, Jan. 2018, pp. 2976–2989.

[75] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent.*, Sep. 2015, pp. 1–15.

[76] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, vol. 4, Feb. 2018, pp. 2587–2601.

[77] R. Yan, Q. Xing, and Y. Xu, "Multi agent safe graph reinforcement learning for PV inverter S based real-time de centralized Volt/Var control in zoned distribution networks," *IEEE Trans. Smart Grid*, vol. 15, no. 1, pp. 299–311, 2023, doi: 10.1109/TSG.2023.3277087.

[78] T. Fan, X. Y. Lee, and Y. Wang, "PowerGym: A reinforcement learning environment for Volt-Var control in power distribution systems," in *Proc. Learn. Dyn. Control Conf.*, Sep. 2021, no. 1, pp. 1–19.

[79] J. G. Vlachogiannis and N. D. Hatziargyriou, "Reinforcement learning for reactive power control," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1317–1325, Aug. 2004, doi: 10.1109/TPWRS.2004.831259.

[80] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 1990–2001, May 2020, doi: 10.1109/TPWRS.2019.2948132.

[81] R. Hossain, M. M. Lakouraj, A. Ghasemkhani, H. Livani, and M. Ben-Idris, "Deep reinforcement learning-based Volt-Var optimization in distribution grids with inverter-based resources," in *Proc. North Amer. Power Symp. (NAPS)*, Nov. 2021, pp. 1–6, doi: 10.1109/NAPS52732.2021.9654630.

[82] R. Hossain, M. Gautam, M. M. Lakouraj, H. Livani, and M. Benidris, "Volt-Var optimization in distribution networks using twin delayed deep reinforcement learning," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Apr. 2022, pp. 1–5, doi: 10.1109/ISGT50606.2022.9817477.

[83] R. Hossain, M. Gautam, M. MansourLakouraj, H. Livani, M. Benidris, and Y. Baghzouz, "Soft actor critic based Volt-Var co-optimization in active distribution grids," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2022, pp. 1–5, doi: 10.1109/PESGM48719.2022.9916976.

[84] B. Zhang, H. Liu, G. Zhuang, L. Liu, and W. Wu, "Data-driven wind farm Volt/Var control based on deep reinforcement learning," in *Proc. IEEE 4th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Oct. 2020, pp. 2758–2763, doi: 10.1109/EI250167.2020.9347074.

[85] W. Li, W. Huang, T. Zhu, M. Wu, and Z. Yan, "Deep reinforcement learning based continuous Volt-Var optimization in power distribution systems with renewable energy resources," in *Proc. IEEE Sustain. Power Energy Conf. (iSPEC)*, Dec. 2021, pp. 682–686, doi: 10.1109/iSPEC53008.2021.9735939.

[86] R. Si, T. Gao, Y. Dai, Y. Bai, Y. Jiang, and J. Zhang, "Evolutionary deep reinforcement learning for Volt-Var control in distribution network," in *Proc. IEEE 2nd Int. Conf. Digit. Twins Parallel Intell. (DTPI)*, Oct. 2022, pp. 1–18, doi: 10.1109/dtpi55838.2022.9998947.

[87] T. Zhao, Y. Zhang, and M. Yue, "Scalable deep reinforcement learning-based Volt-Var optimization in distribution systems: A mean-field approach," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2022, pp. 1–5, doi: 10.1109/PESGM48719.2022.9916943.

[88] X. Y. Lee, S. Sarkar, and Y. Wang, "A graph policy network approach for Volt-Var control in power distribution systems," *Appl. Energy*, vol. 323, Oct. 2022, Art. no. 119530, doi: 10.1016/j.apenergy.2022.119530.

[89] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2313–2323, May 2020, doi: 10.1109/TSG.2019.2951769.

[90] F. Kabir, Y. Gao, and N. Yu, "Reinforcement learning-based smart inverter control with polar action space in power distribution systems," in *Proc. IEEE Conf. Control Technol. Appl. (CCTA)*, Aug. 2021, pp. 315–322, doi: 10.1109/CCTA48906.2021.9659162.

[91] H. Wu, Z. Xu, M. Wang, J. Zhao, and X. Xu, "Two-stage voltage regulation in power distribution system using graph convolutional network-based deep reinforcement learning in real time," *Int. J. Electr. Power Energy Syst.*, vol. 151, Sep. 2023, Art. no. 109158, doi: 10.1016/j.ijepes.2023.109158.

[92] R. Hossain, M. Gautam, J. Thapa, H. Livani, and M. Benidris, "Deep reinforcement learning assisted co-optimization of Volt-Var grid service in distribution networks," *Sustain. Energy, Grids Netw.*, vol. 35, Sep. 2023, Art. no. 101086, doi: 10.1016/j.segan.2023.101086.

[93] D. Cao, J. Zhao, W. Hu, F. Ding, N. Yu, Q. Huang, and Z. Chen, "Model-free voltage control of active distribution system with PVs using surrogate model-based deep reinforcement learning," *Appl. Energy*, vol. 306, Jan. 2022, Art. no. 117982, doi: 10.1016/j.apenergy.2021.117982.

[94] D. Cao, J. Zhao, J. Hu, Y. Pei, Q. Huang, Z. Chen, and W. Hu, "Physics-informed graphical representation-enabled deep reinforcement learning for robust distribution system voltage control," *IEEE Trans. Smart Grid*, vol. 15, no. 1, pp. 233–246, 2023, doi: 10.1109/TSG.2023.3267069.

[95] A. Pigott, C. Crozier, K. Baker, and Z. Nagy, "GridLearn: Multiagent reinforcement learning for grid-aware building energy management," *Electric Power Syst. Res.*, vol. 213, Dec. 2022, Art. no. 108521, doi: 10.1016/j.epsr.2022.108521.

[96] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 4120–4123, Sep. 2020, doi: 10.1109/TPWRS.2020.3000652.

[97] S. Zhang, M. Zhang, R. Hu, D. Lubkeman, Y. Liu, and N. Lu, "Reinforcement learning for Volt-Var control: A novel two-stage progressive training strategy," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2022, pp. 1–5, doi: 10.1109/PESGM48719.2022.9916659.

[98] Y. Lu, Y. Xiang, Y. Huang, B. Yu, L. Weng, and J. Liu, "Deep reinforcement learning based optimal scheduling of active distribution system considering distributed generation, energy storage and flexible load," *Energy*, vol. 271, May 2023, Art. no. 127087, doi: 10.1016/j.energy.2023.127087.

[99] Y. Pei, Y. Yao, J. Zhao, F. Ding, and K. Ye, "Data-driven distribution system coordinated PV inverter control using deep reinforcement learning," in *Proc. IEEE Sustain. Power Energy Conf. (iSPEC)*, Dec. 2021, pp. 781–786, doi: 10.1109/iSPEC53008.2021.9735897.

[100] D. Hu, Y. Peng, J. Yang, Q. Deng, and T. Cai, "Deep reinforcement learning based coordinated voltage control in smart distribution network," in *Proc. Int. Conf. Power Syst. Technol.*, Dec. 2021, pp. 1030–1034, doi: 10.1109/POWERCON53785.2021.9697762.

[101] T. Zhu, G. Lu, Y. Duan, D. Hai, S. Zhou, R. Zhang, and J. Wei, "SAC-based multi-agent framework for continuous Volt-Var control in distribution network with high penetration of PVs," in *Proc. IEEE 5th Int. Electr. Energy Conf. (CIEEC)*, May 2022, pp. 4831–4835, doi: 10.1109/CIEEC54735.2022.9845852.

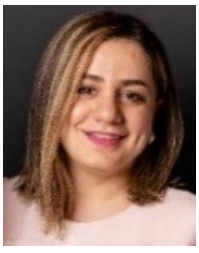
[102] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, and Z. Chen, "Attention enabled multi-agent DRL for decentralized Volt-Var control of active distribution system using PV inverters and SVCs," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1582–1592, Jul. 2021, doi: 10.1109/TSTE.2021.3057090.

[103] Q. Liu, Y. Guo, L. Deng, H. Liu, D. Li, H. Sun, and W. Huang, "Reducing learning difficulties: One-step two-critic deep reinforcement learning for inverter-based Volt-Var control," 2022, *arxiv:2203.16289*.

[104] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based Volt-Var optimization in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361–371, Jan. 2021, doi: 10.1109/TSG.2020.3010130.

[105] Y. Gao, W. Wang, and N. Yu, "Consensus multi-agent reinforcement learning for Volt-Var control in power distribution networks," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 3594–3604, Jul. 2021, doi: 10.1109/TSG.2021.3058996.

[106] H. Li, Z. Wang, and H. He, "Distributed Volt-Var optimization based on multi-agent deep reinforcement learning," in *Proc. Int. Joint Conf. Neural Networks*, 2021, pp. 1–17.
[107] B. Wang, Y. Xu, S. B. Hee, and Z. Yan, "A multi-agent deep reinforcement learning based multi-timescale voltage control for distribution system," in *Proc. IEEE 5th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Oct. 2021, pp. 2825–2830, doi: 10.1109/EI252483.2021.9713354.
[108] F. Kabir, N. Yu, Y. Gao, and W. Wang, "Deep reinforcement learning-based two-timescale Volt-Var control with degradation-aware smart inverters in power distribution systems," *Appl. Energy*, vol. 335, Apr. 2023, Art. no. 120629, doi: 10.1016/j.apenergy.2022.120629.
[109] D. Cao, J. Zhao, W. Hu, N. Yu, F. Ding, Q. Huang, and Z. Chen, "Deep reinforcement learning enabled physical-model-free two-timescale voltage control method for active distribution systems," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 149–165, Jan. 2022, doi: 10.1109/TSG.2021.3113085.
[110] H. Liu, W. Wu, and Y. Wang, "Bi-level off-policy reinforcement learning for two-timescale Volt/Var control in active distribution networks," *IEEE Trans. Power Syst.*, vol. 38, no. 1, pp. 385–395, Jan. 2023, doi: 10.1109/TPWRS.2022.3168700.
[111] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for Volt-Var control in power distribution systems," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3008–3018, Jul. 2020, doi: 10.1109/TSG.2019.2962625.
[112] W. Wang, N. Yu, J. Shi, and Y. Gao, "Volt-Var control in power distribution systems with deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids*, Oct. 2019, pp. 1–7, doi: 10.1109/SMARTGRIDCOMM.2019.8909741.
[113] Y. Gao and N. Yu, "Model-augmented safe reinforcement learning for Volt-Var control in power distribution networks," *Appl. Energy*, vol. 313, May 2022, Art. no. 118762, doi: 10.1016/j.apenergy.2022.118762.
[114] P. Kou, D. Liang, C. Wang, Z. Wu, and L. Gao, "Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks," *Appl. Energy*, vol. 264, Apr. 2020, Art. no. 114772, doi: 10.1016/j.apenergy.2020.114772.
[115] H. Li and H. He, "Learning to operate distribution networks with safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 13, no. 3, pp. 1860–1872, May 2022, doi: 10.1109/TSG.2022.3142961.
[116] H. T. Nguyen and D.-H. Choi, "Three-stage inverter-based peak shaving and Volt-Var control in active distribution networks using online safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 3266–3277, Jul. 2022, doi: 10.1109/TSG.2022.3166192.
[117] H. Liu and W. Wu, "Online multi-agent reinforcement learning for decentralized inverter-based Volt-Var control," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 2980–2990, Jul. 2021, doi: 10.1109/TSG.2021.3060027.
[118] P. Chen, S. Liu, X. Wang, and I. Kamwa, "Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems," *IEEE Trans. Smart Grid*, vol. 14, no. 4, pp. 2656–2667, 2022, doi: 10.1109/TSG.2022.3228636.
[119] S. Afrasiabi, M. Afrasiabi, M. A. Jarrahi, M. Mohammadi, J. Aghaei, M. S. Javadi, M. Shafie-Khah, and J. P. S. Catalão, "Wide-Area composite load parameter identification based on multi-residual deep neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 1–11, Sep. 2021, doi: 10.1109/TNNLS.2021.3133350.



SARAH ALLAHMORADI (Member, IEEE) received the B.Sc. degree from the University of Kurdistan, Kurdistan, Iran, in 2016, and the M.Sc. degree (Hons.) from Tarbiat Modares University, Tehran, Iran, in 2020, both in electrical engineering. She is currently pursuing the Ph.D. degree with the University of Saskatchewan, Saskatoon, Canada. Her research interests include smart grids, optimization, and machine learning applications in power systems.

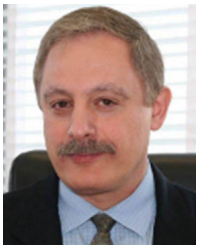
SHAHABODIN AFRASIABI (Member, IEEE) received the B.Sc. degree from Semenan University, Semnan, Iran, in 2014, and the M.Sc. degree from Shahid Chamran University, Ahvaz, Iran, in 2017, both in electrical engineering. He is currently pursuing the Ph.D. degree with the University of Saskatchewan, Saskatoon, Canada. His research interests include power system dynamics, machine learning, state estimation, and power system probabilistic analysis.

XIAODONG LIANG (Senior Member, IEEE) was born in Lingyuan, Liaoning, China. She received the B.Eng. and M.Eng. degrees from Shenyang Polytechnic University, Shenyang, China, in 1992 and 1995, respectively, the M.Sc. degree from the University of Saskatchewan, Saskatoon, Canada, in 2004, and the Ph.D. degree from the University of Alberta, Edmonton, Canada, in 2013, all in electrical engineering.

From 1995 to 1999, she was a Lecturer with Northeastern University, Shenyang, China. In October 2001, she joined Schlumberger (SLB), Edmonton, Canada, and was promoted to a Principal Power Systems Engineer with this world's leading oil field service company, in 2009. She was with Schlumberger for almost 12 years until August 2013. From 2013 to 2019, she was with Washington State University, Vancouver, WA, USA, and the Memorial University of Newfoundland, in St. John's, NL, Canada, as an Assistant Professor and later an Associate Professor. In July 2019, she joined the University of Saskatchewan, Saskatoon, Canada, where she is currently a Professor and the Canada Research Chair in technology solutions for energy security in remote, northern, and indigenous communities. She was an Adjunct Professor with the Memorial University of Newfoundland, from 2019 to 2022. Her research interests include power systems, renewable energy, and electric machines.

Dr. Liang is fellow of IET. She is the registered Professional Engineer in the province of Saskatchewan, Canada; and the Deputy Editor-in-Chief of IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS.

JUNBO ZHAO (Senior Member, IEEE) is the Associate Director of Eversource Energy Center for Grid Modernization and Strategic Partnerships; and an Assistant Professor with the Department of Electrical and Computer Engineering, University of Connecticut. He is also a Research Scientist with the National Renewable Energy Laboratory. He serves as an Associate Editor for IEEE TRANSACTIONS ON POWER SYSTEMS and IEEE TRANSACTIONS ON SMART GRID, a North America Regional Editor for *IET Renewable Power Generation*, and a Subject Editor for *IET Generation, Transmission and Distribution*.



MOHAMMAD SHAHIDEHPOUR (Life Fellow, IEEE) received the Honorary Doctorate degree in electrical engineering from the Politehnica University of Bucharest, Bucharest, Romania. He is currently a University Distinguished Professor with Illinois Institute of Technology, Chicago, IL, USA, where he is also the Bodine Chair Professor and the Director of the Robert W. Galvin Center for Electricity Innovation. He is a member of the U.S. National Academy of Engineering and a fellow of the American Association for the Advancement of Science and the National Academy of Inventors.