DistFlow Safe Reinforcement Learning Algorithm for Voltage Magnitude Regulation in Distribution Networks

Shengren Hou, *Student Member*, *IEEE*, Aihui Fu, *Member*, *IEEE*, Edgar Mauricio Salazar Duque, *Member*, *IEEE*, Peter Palensky, *Senior Member*, *IEEE*, Qixin Chen, *Senior Member*, *IEEE*, and Pedro P. Vergara, *Senior Member*, *IEEE*

***Abstract*—The integration of distributed energy resources (DERs) has escalated the challenge of voltage magnitude regulation in distribution networks. Model-based approaches, which rely on complex sequential mathematical formulations, cannot meet the real-time demand. Deep reinforcement learning (DRL) offers an alternative by utilizing offline training with distribution network simulators and then executing online without computation. However, DRL algorithms fail to enforce voltage magnitude constraints during training and testing, potentially leading to serious operational violations. To tackle these challenges, we introduce a novel safe-guaranteed reinforcement learning algorithm, the DistFlow safe reinforcement learning (DF-SRL), designed specifically for real-time voltage magnitude regulation in distribution networks. The DF-SRL algorithm incorporates a DistFlow linearization to construct an expert-knowledge-based safety layer. Subsequently, the DF-SRL algorithm overlays this safety layer on top of the agent policy, recalibrating unsafe actions to safe domains through a quadratic programming formulation. Simulation results show the DF-SRL algorithm consistently ensures voltage magnitude constraints during training and real-time operation (test) phases, achieving faster convergence and higher performance, which differentiates it apart from (safe) DRL benchmark algorithms.**

***Index Terms*—Voltage regulation, distribution network, safe reinforcement learning, energy management.**

NOMENCLATURE

A. Sets and Indexes

B	Batch of data collected from reply buffer to update policy and critic models
$\mathcal{L}$	Set of lines connecting nodes in distribution network
m, n	Node indexes
$\mathcal{N}$	Set of nodes in distribution network
t	Time step index
$\mathcal{T}$	Set of time steps

B. Parameters

σ_1	Standard deviation of Gaussian distribution
θ, ∇_θ	Parameters of critic network and related gradient
ω, ∇_ω	Parameters of trained policy and related gradient
ϵ	Small value added to control relaxation condition of voltage magnitude limits
τ	Iterative number
$F_{l,i}$, $T_{l,j}$	Elements of connection matrices $\boldsymbol{F}$ and $\boldsymbol{T}$
$f(\cdot)$, $t(\cdot)$	"From" and "to" node indexes for lines
$\bar{i}^2_{mn}$	The maximum current limit of line connecting node m and node n
$N(\cdot)$	Gaussian distribution
$p^S_{m,t}$, $q^S_{m,t}$	Active and reactive power demands at node m and time step t
$p^{EV}_{m,t}$	Active power from electric vehicles (EVs) at node m and time step t
$\bar{p}^B_m$, $\underline{p}^B_m$	The maximum and minimum active power provided by aggregator at node m
$\bar{p}^B_{m,t}$, $\underline{p}^B_{m,t}$	The maximum and minimum active power provided by aggregator at node m and time step t
$p^D_{m,t}$, $q^D_{m,t}$	Active and reactive power demands at node m and time step t
$p^{PV}_{m,t}$	Active power generation of photovoltaic (PV) systems at node m and time step t

Manuscript received: March 8, 2024; revised: April 28, 2024; accepted: July 17, 2024. Date of CrossCheck: July 17, 2024. Date of online publication: August 26, 2024.

This work is part of the DATALESs project (with project number 482.20.602) jointly financed by the Netherlands Organization for Scientific Research (NWO) and the National Natural Science Foundation of China (NSFC).

S. Hou, A. Fu, P. Palensky, and P. P. Vergara (corresponding author) are with the Intelligent Electrical Power Grids (IEPG) Group, Delft University of Technology, Delft 2628CD, The Netherlands (e-mail: h.shengren@tudeflt.nl; A.fu@tudeflt.nl; P.Palensky@tudeflt.nl; P.P.VergaraBarrios@tudeflt.nl).

E. M. S. Duque is with the Electrical Energy Systems (EES) Group, Eindhoven University of Technology, Eindhoven, The Netherlands (e-mail: e.m.salazar.duque@tue.nl).

Q. Chen is with the State Key Laboratory of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China (e-mail: qxchen@tsinghua.edu.cn).

DOI: 10.35833/MPCE.2024.000253

$p^{\mathrm{N}}_{m,t}$ Net active power of node m at time step t

Q_{θ_i}, $Q_{\theta_i^{\mathrm{target}}}$ Trained i^{th} Q-function and target Q-function

r_{mn}, x_{mn} Resistance and reactance of line connecting nodes m and n

$\bar{v}$, $\underline{v}$ The maximum and minimum voltage magnitude limits

$\bar{v}^2$, $\underline{v}^2$ Upper and lower bounds for squared voltage magnitude

v_0 Voltage magnitude at slack node, which is typically considered constant and known

C. Continuous Variables

$i_{mn,t}$ Current magnitude in line connecting nodes m and n at time step t

$p^B_{m,t}$ Flexible active power provided by aggregator at node m and time step t

$p_{mn,t}$, $q_{mn,t}$ Active and reactive power flows from node m to node n at time step t

p^S_t, q^S_t Active and reactive power injections of slack node at time step t

p^B_i Active power flexibility provided at node i

p_i, q_i Active and reactive power of node i

v_i Voltage magnitude of node i

$v_{m,t}$ Voltage magnitude of node m at time step t

D. Matrices and Vectors

$\boldsymbol{a}$, $\hat{\boldsymbol{a}}$ Original and projected (safe) action vectors

$\boldsymbol{B}$, $\boldsymbol{C}$ Matrices used in linear power flow formulation

$\boldsymbol{D}(\boldsymbol{r}_{mn})$, $\boldsymbol{D}(\boldsymbol{x}_{mn})$ Diagonal matrices constructed from $\boldsymbol{r}_{mn}$ and $\boldsymbol{x}_{mn}$

$\boldsymbol{F}$, $\boldsymbol{T}$ Connection matrices representing "from" and "to" nodes of lines

$\boldsymbol{I}$ Unit matrix

$\boldsymbol{M}$ Full incident matrix of distribution network

$\boldsymbol{M}_0$ Incidence matrix of distribution network

$\boldsymbol{m}_0$ Column of incidence matrix corresponding to slack node

$\boldsymbol{p}^{\mathrm{N}}_m$, $\boldsymbol{q}^{\mathrm{N}}_m$ Vectors representing net active and reactive power injections

$\boldsymbol{r}_{mn}$, $\boldsymbol{x}_{mn}$ Vectors representing resistance and reactance of lines

$\boldsymbol{v}^2$ Vector representing squared voltage magnitude of nodes

$\boldsymbol{v}_m$ Vector representing voltage magnitude of node m

$\mathbf{1}_{|\mathcal{L}|}$ Unit vector with dimension equal to number of lines in network

I. Introduction

DISTRIBUTION networks have experienced a notable increase in distributed energy resource (DER) integration, including residential photovoltaic (PV) systems, energy storage systems (ESSs), and plug-in electric vehicles (EVs) [1], [2]. This rise in DERs contributes to sustainability efforts and poses operational challenges to distribution system operators (DSOs). Among these challenges, the voltage magnitude regulation has surfaced as a predominant concern [3]. Aggregators, who control various DERs, have stepped in to offer a solution. By providing significant flexibility to DSOs, aggregators enable the strategic procurement and deployment of this flexibility, thereby facilitating efficient voltage magnitude regulation [4].

Implementing voltage magnitude regulation adopts one of two approaches: model-based and model-free approaches. Model-based approaches manage voltage magnitude regulation by solving mathematical formulations defined via an objective function and a set of operational constraints [5]. However, the intricacy of these model-based approaches increases with the complexity of distribution networks and sequential regulation slots because they necessitate complete network and DER information. Therefore, solving such formulations can be computationally intensive and thus cannot meet real-time demand [6]. Conversely, the model-free deep reinforcement learning (DRL) represents an alternative approach that does not require online computation by leveraging an offline training procedure and distribution network simulators [7]. Nevertheless, a significant drawback of such DRL algorithms is their inability to ensure action feasibility and, thus, safety [8], [9]. To address this, some studies have formulated the voltage magnitude constraint as a soft constraint, i.e., a fixed [10], [11] or trainable penalty term [12], which is added to the reward function and used to guide the DRL algorithm during training. For instance, the reinforcement learning (RL) algorithm proposed in [13] follows this approach, which is developed to define the ESS schedule to minimize operational costs while respecting voltage magnitude limits. Nevertheless, this approach fails to enforce such constraints strictly during training and real-time operation.

Several safe DRL algorithms have recently been developed to enforce operational constraints in control systems [14]. In [15], a constrained soft actor-critic (SAC) algorithm was developed for EV charging in residential microgrids to cater to the increasing prominence of EVs. Using a constrained Markov decision process (MDP) formulation and a ladder electricity pricing scheme, this algorithm showed promising results in reducing action space dimensionality and ensuring safe EV charging. Another study [16] implemented primal-dual optimization within a safe RL framework, showing superior performance in terms of energy cost minimization and constraint adherence. In [17], a safe DRL algorithm was introduced to define a fast-charging strategy for lithium-ion ESSs to enhance the efficiency of EV charging without compromising ESS safety. Utilizing the SAC-Lagrange DRL within a cyber-physical system framework, this algorithm optimizes charging speeds by leveraging an electro-thermal model, outperforming existing deep deterministic policy gradient (DDPG) based and SAC-based DRL algorithms in terms of optimality.

To ensure that the updated policy stays within a feasible set, a cumulative constraint violation index was kept below a predetermined threshold in [18] and [19]. This approach was also used in [20] and [21], in which the constraint viola-

tion index is designed to reflect the voltage and current magnitude violation levels due to the ESS dispatch defined. Nevertheless, this constraint policy was initially developed to handle cumulative or chance constraints after training [19]. On the contrary, voltage magnitude violation issues in distribution networks are state-wise constraints, which do not rely on historical trajectories or random variables but hinge on the current state of the environment [22]. Consequently, applying constraint policy optimization methods to voltage regulation issues cannot offer a probabilistic sense of safety. In [23] and [24], the trained DRL algorithm was formulated as a mixed-integer programming (MIP) formulation and voltage magnitude constraints were added to the MIP. By solving this extended MIP, the actions from the DRL algorithm are projected into safe action spaces that strictly enforce constraints. Nevertheless, this approach cannot meet the real-time operation requirements if the formulated MIP becomes too large. In [25], the stability of distribution network controlled by DRL algorithms was guaranteed if the system adheres to specific Lipschitz constraints. However, formulating such Lipschitz sets for distribution networks is quite changeable. In [26], a constrained SAC algorithm was proposed to address volt-var control challenges. The constrained SAC algorithm combines the maximum entropy framework, the method of multiplier, a device-decoupled neural network structure, and an ordinal encoding scheme to achieve scalability, sample efficiency, and constraint satisfaction. However, the algorithm can only be applied to discrete action problems.

Safety layer-based DRL algorithms are suitable to handle the state-wise constraints (i.e., voltage magnitude), which formulate a policy-independent safety layer to project actions defined by DRL algorithms into a feasible set. In [27], a deep neural network (DNN) assisted projection-based DRL algorithm was proposed for the safe control of distribution networks. This algorithm leverages a pre-trained DNN to accelerate the projection calculations, enabling the rapid identification of safe actions. However, a critical limitation of this algorithm is the reliability of the safe actions produced by the DNN. Since the DNN is trained on historical data, the quality and representativeness of the data are paramount. Alternatively, a linear safety layer is trained by the data collected from a random policy with the environment [28]. In [29], a safety layer was built upon DRL algorithms to filter out unsafe actions before the execution, while voltage magnitude was enforced by solving projection. A similar approach was implemented in [30] to regulate the voltage magnitude of the distribution network via controlling smart transformers. Yet, these algorithms mainly rely on training a linear safety layer to first capture the sensitivity between station-action pair and constraint violations, and then filter out unsafe actions before they are executed. Therefore, the safety guarantee performance for these algorithms is highly dependent on the quality of the trained linear safety layer. Given the complex relationships between system dynamics and multi-dimension constraints involved in voltage magnitude regulation problems, training such a linear safety layer often proves to be a significant challenge [28]. Consequently, the trained safety layer can rarely provide a safety guarantee for the voltage magnitude regulation problem in the distribution network, leading to sub-optimal performance and violations.

Drawing on the pivotal insights [31]-[33] that integrating expert knowledge can significantly enhance safety and agent performance, we introduce the DistFlow safe reinforcement learning (DF-SRL) algorithm. It aims to tackle state-wise voltage magnitude regulation issues in distribution networks by applying DRL algorithms, augmented with an expert-knowledge-based safety layer. This innovation addresses existing gaps in voltage magnitude regulation research through several key contributions:

1) The proposed DF-SRL algorithm incorporates a DistFlow linearization to devise a safety layer, leveraging expert knowledge insights to accurately map the relationship between actions of the agent and voltage magnitude variations in distribution networks.

2) The DRL algorithm overlays the safety layer on top of the DRL policy to recalibrate potentially unsafe actions to conform to safe parameters by optimizing the proximity of these actions in Euclidean space.

3) The error of the safety layer introduced by linearization is corrected by the slack parameter, and a detailed sensitivity and scalability analysis is conducted.

4) The proposed DF-SRL algorithm ensures the practicality and real-time viability of actions and guarantees safety constraints during both the training and application phases.

II. Voltage Magnitude Regulation Problem

Voltage fluctuations in distribution networks are predominantly due to variations in active power, such as those caused by overload conditions or high inflows from PV systems [3]. These fluctuations are more directly linked to active power changes, affecting voltage magnitude significantly. By focusing on active power, aggregators can utilize DERs like battery storage and controllable loads more effectively. This aligns with operational strategies that maximize the impact of available resources while ensuring compliance with safety and reliability standards.

The voltage magnitude regulation framework for DSO and aggregators is depicted in Fig. 1.

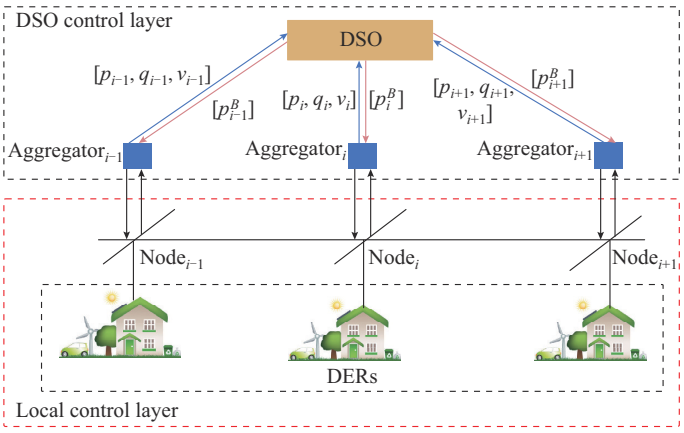

Fig. 1. Voltage magnitude regulation framework for DSO and aggregators.

Each network node is associated with an aggregator that oversees a group of consumers equipped with DERs such as residential PV systems, ESSs, and plug-in EVs. These aggregators are empowered to fully control the DERs of their designated consumers, playing a pivotal role in the dynamic management of the distribution network. Aggregators collect consumer data, build baseline electrical consumption profiles, and share the active power flexibility with the DSO control center. Subsequently, the DSO control center deploys a voltage magnitude regulation algorithm to determine the required active power flexibility that each aggregator must provide.

In this paper, we focus on developing an RL-based algorithm to assist the DSO control center in accurately determining the required flexibility provision of each aggregator to achieve voltage magnitude regulation.

A. Mathematical Programming Formulation

In general, the voltage magnitude regulation problem can be modeled using the non-linear programming (NLP) formulation given by (1)-(9). The objective function in (1) aims to minimize the use of flexible active power $p^B_{m,t}$ provided by all aggregators within the set $m \in \mathcal{N}$, aiming to regulate the voltage magnitude over the time horizon $\mathcal{T}$.

$$\min_{p^B_{m,t}} \left\{ \sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{N}} |p^B_{m,t}| \Delta t \right\} \tag{1}$$

s.t.

$$\sum_{nm \in \mathcal{L}} p_{nm,t} - \sum_{mn \in \mathcal{L}} \left(p_{mn,t} + r_{mn} i^2_{mn,t} \right) + p^B_{m,t} + p^{PV}_{m,t} + p^S_{m,t} = p^D_{m,t} + p^{EV}_{m,t} \quad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{2}$$

$$\sum_{nm \in \mathcal{L}} q_{nm,t} - \sum_{mn \in \mathcal{L}} \left(q_{mn,t} + x_{mn} i^2_{mn,t} \right) + q^S_{m,t} = q^D_{m,t} \quad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{3}$$

$$v^2_{m,t} - v^2_{n,t} = 2\left(r_{mn} p_{mn,t} + x_{mn} q_{mn,t} \right) + \left(r^2_{mn} + x^2_{mn} \right) i^2_{mn,t} \quad \forall m,n \in \mathcal{N}, \forall t \in \mathcal{T} \tag{4}$$

$$v^2_{m,t} i^2_{mn,t} = p^2_{mn,t} + q^2_{mn,t} \quad \forall m,n \in \mathcal{N}, \forall t \in \mathcal{T} \tag{5}$$

$$\underline{p}^B_{m,t} \le p^B_{m,t} \le \bar{p}^B_{m,t} \quad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{6}$$

$$\underline{v}^2 \le v^2_{m,t} \le \bar{v}^2 \quad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{7}$$

$$0 \le i^2_{mn,t} \le \bar{i}^2_{mn} \quad \forall mn \in \mathcal{L}, \forall t \in \mathcal{T} \tag{8}$$

$$p^S_{m,t} = q^S_{m,t} = 0 \quad \forall m \in \mathcal{N} \setminus \{1\}, \forall t \in \mathcal{T} \tag{9}$$

The distribution network is formulated based on the power flow formulation shown in (2)-(5), according to the active power $p_{mn,t}$, reactive power $q_{mn,t}$, and current magnitude $i_{mn,t}$ of lines, and the voltage magnitude $v_{m,t}$ of nodes. The expression in (6) enforces the used flexible active power within the boundaries that each aggregator provides, while (7) and (8) enforce the voltage magnitude and line current limits, respectively. Finally, (9) enforces that only one node is connected to the substation. Flexibility for voltage magnitude regulation at each aggregator can vary over day and time slots [3].

B. Constrained Markov Decision Process (CMDP) Formulation

The voltage magnitude regulation problem can be modeled as a case of CMDPs, represented by a 6-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mathcal{C})$. Here, $\mathcal{S}$ denotes a state space encompassing the observable states of the distribution network; $\mathcal{A}$ denotes an action space representing the possible control actions; $\mathcal{P}$ is the state transition probability function capturing the system dynamics; $\mathcal{R}$ is the reward function guiding the optimization; γ is a discount factor reflecting the importance of future rewards; and $\mathcal{C}$ is a set of immediate constraint functions ensuring operational safety and feasibility. The decision as to which action a_t is chosen in a certain state s_t is governed by a policy $\pi(a_t|s_t)$. The agent employs the policy to interact with the formulated CMDP and define a trajectory of states, actions, and rewards: $\tau = (s_0, a_0, s_1, a_1, \ldots)$. This trajectory not only aims to maximize the cumulative reward but also adheres to the system constraints, thereby balancing the objectives of operational efficiency and safety.

1) State

The state at time t encapsulates the current operational status of the distribution network, providing a comprehensive view of the system dynamics, and it is defined by:

$$s_t = \left(p^{\mathrm{N}}_{m,t}, v_{m,t}, \underline{p}^B_{m,t}, \bar{p}^B_{m,t} \right) \quad m \in \mathcal{N} \tag{10}$$

where $p^{\mathrm{N}}_{m,t} = p^D_{m,t} - p^{PV}_{m,t} - p^{EV}_{m,t}$, which captures the balance among the demand, PV generation, and EV consumption at node m.

2) Action

The action space $\mathcal{A}$ consists of the set of all possible active power adjustments at each node m, defined as $\mathcal{A} = \left\{ a_t \middle| a_t = p^B_{m,t}, \underline{p}^B_{m,t} \le p^B_{m,t} \le \bar{p}^B_{m,t}, \forall m \in \mathcal{N} \right\}$.

3) Reward

The DSO seeks to regulate the voltage magnitude into defined boundaries while minimizing the use of total active power flexibility provided by aggregators. Thus, the reward function r_t is defined as the negative of the total used flexible active power, which can be expressed as:

$$r_t = -\sum_{m \in \mathcal{N}} \left| p^B_{m,t} \right| \tag{11}$$

This formulation incentivizes the minimization of the total active power flexibility utilized, thereby promoting energy efficiency and cost effectiveness in voltage magnitude regulation. Given the state s_t and action a_t at time step t, the system transits to the next state s_{t+1} defined by the transition probability function that can be expressed as:

$$p\left(S_{t+1}, R_t \middle| S_t, A_t \right) = \Pr\left\{ S_{t+1} = s_{t+1}, R_t = r_t \middle| S_t = s_t, A_t = a_t \right\} \tag{12}$$

where $R_t | S_t, A_t$ is the reward distribution under the current state S_t and action A_t. The goal of the RL agent is to find a policy that maximizes the cumulative discounted return $J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t r_t \right]$ while ensuring no constraint is violated during the exploration and exploitation processes. $\mathbb{E}_{\tau \sim \pi}[\cdot]$ is the expectation function of the trajectory distribution under

the current policy. $\sum_{t=0}^{T}\gamma^t r_t$ is the cumulative return in current trajectory. The penalty term induced by the constraint violations $C_{m,t}(\pi)$ denotes the voltage magnitude violation of node m at time step t, which is defined as:

$$C_{m,t}(\pi)=\max\left\{0,\left|v_0-v_{m,t}\right|-\frac{\bar{v}-\underline{v}}{2}\right\}\quad\forall m\in\mathcal{N}\tag{13}$$

This formulation ensures that $C_{m,t}(\pi)$ represents a positive penalty term when the voltage magnitude at node m deviates outside the acceptable range defined by $\underline{v}$ and $\bar{v}$, and is zero otherwise.

The voltage magnitude regulation problem formulated as a CMDP can then be expressed using the following constrained optimization formulation:

$$\begin{cases}\max\limits_{\pi}=\mathbb{E}_{\tau\sim\pi}\left[\sum\limits_{t=0}^{\mathcal{T}}\gamma^t r_t\right]\\ \text{s.t. } C_{m,t}(\pi)=0\quad\forall m\in\mathcal{N},\forall t\in\mathcal{T}\end{cases}\tag{14}$$

In this formulation, $C_{m,t}(\pi)$ serves as a constraint in the CMDP, ensuring that the policy π leads to actions that maintain the voltage magnitude within the specified limits. It is indirectly influenced by the policy through its impact on the state s_t and the action a_t.

III. Proposed DF-SRL Algorithm

The proposed DF-SRL algorithm is defined through a parameterized policy network, denoted by $\pi_\omega(\cdot)$. This policy network selects actions based on the current state, performing exploration and exploitation. To enhance safety and ensure that voltage magnitude constraints are met during the exploration, we introduce a safety layer on top of the policy network $\pi_\omega(\cdot)$. A safety layer is designed based on the parameters and topology of the distribution network, enabling a projection of the original action proposed by the RL algorithm onto a safe domain. A more detailed explanation is provided as follows.

A. DRL Algorithms

Traditional value-based DRL algorithms fail to solve the voltage magnitude regulation problem due to the continuous nature of the state and action spaces [34]. Alternatively, policy-based DRL algorithms such as DDPG [35] and twin delayed deep deterministic policy gradient (TD3) [36] are capable of handling continuous actions by simultaneously maintaining a policy (actor) network $\pi_\omega(s_t)$ that is used to sample actions and a trained Q-function (critic) $Q_\theta(s_t,a_t)$ that is used to guide the update direction of the policy network. The TD3 algorithm is an improved version of the DDPG algorithm, which uses two Q-networks and delayed critic network improvement to reduce the overestimation bias of the critic network in DDPG algorithm. In general, the TD3 algorithm updates the actor network as (15), while the critic update iteration is defined as (16).

$$\omega\leftarrow\omega+\nabla_\omega\frac{1}{|B|}\sum_{s_t\in B}\left(\min_{i=1,2}\left\{Q_{\theta_i}\left(s_t,\pi_\omega(s_t)\right)\right\}\right)\tag{15}$$

$$\min_{\theta}\sum_{s\in B}\left(r_t+\gamma\min_{i=1,2}\left\{Q_{\theta_i^{\text{target}}}\left(s_{t+1},\pi_\omega(s_{t+1})\right)\right\}-Q_{\theta_i}(s_t,a_t)\right)^2\tag{16}$$

Although the TD3 algorithm effectively handles continuous action space problems, it cannot enforce constraints during the training and testing. To solve the CMDP formulation using the TD3 algorithm, the constraint violations $C_{m,t}$ should be added as penalty term to the reward function in (11), defined as:

$$r_t=-\sum_{m\in\mathcal{N}}\left|p_{m,t}^B\right|-\sigma\sum_{m\in\mathcal{N}}C_{m,t}\tag{17}$$

where σ is used to balance the weights between the total required flexibility and the penalty incurred by the voltage magnitude violations. The constrained optimization problem is reformulated into an unconstrained one in this procedure. However, directly applying penalty terms to the reward function cannot guarantee the feasibility strictly, leading to infeasible operations and poor performance [8]. To overcome this, we introduce a linear safety layer on top of the TD3 algorithm to ensure the feasibility of committed actions during the training and testing procedures, as explained in the next subsection.

B. Linear Power Flow Formulation

Given the topology of a distribution network, the incidence matrix $\boldsymbol{M}_0$ can be defined by:

$$\boldsymbol{M}_0=\boldsymbol{F}-\boldsymbol{T}=\left[\boldsymbol{m}_0,\boldsymbol{M}\right]\tag{18}$$

$$F_{\ell,i}=\begin{cases}1 & f(\ell)=i,\ell\in\mathcal{L},i\in\mathcal{N}\\ 0 & \text{otherwise}\end{cases}\tag{19}$$

$$T_{\ell,j}=\begin{cases}1 & t(\ell)=j,\ell\in\mathcal{L},j\in\mathcal{N}\\ 0 & \text{otherwise}\end{cases}\tag{20}$$

Given the diagonal matrices $\boldsymbol{D}(\boldsymbol{r}_{mn})$ and $D(\boldsymbol{x}_{mn})$, the relationship between the voltage magnitude of nodes $\boldsymbol{v}_m$ and the net active and reactive power injections $\boldsymbol{p}_m^{\text{N}}$ and $\boldsymbol{q}_m^{\text{N}}$ can be expressed as:

$$\boldsymbol{M}\boldsymbol{v}_m^2=\boldsymbol{M}v_0^2\boldsymbol{1}_{|\mathcal{L}|}+2\left(\boldsymbol{D}(\boldsymbol{r}_{mn})\boldsymbol{B}\boldsymbol{T}\boldsymbol{p}_m^{\text{N}}+\boldsymbol{D}(\boldsymbol{x}_{mn})\boldsymbol{B}\boldsymbol{T}\boldsymbol{q}_m^{\text{N}}\right)+\boldsymbol{C}\boldsymbol{c}^2\tag{21}$$

$$\boldsymbol{B}=\left(\boldsymbol{I}-\boldsymbol{T}\boldsymbol{F}^{\text{T}}\right)^{-1}\tag{22}$$

$$\boldsymbol{C}=2\left(\boldsymbol{D}(\boldsymbol{r}_{mn})\boldsymbol{B}\boldsymbol{D}(\boldsymbol{r}_{mn})+\boldsymbol{D}(\boldsymbol{x}_{mn})\boldsymbol{B}\boldsymbol{D}(\boldsymbol{x}_{mn})\right)-\boldsymbol{D}\left(\boldsymbol{r}_{mn}^2+\boldsymbol{x}_{mn}^2\right)\tag{23}$$

The linear power flow formulation presented in (21) involves an approximation that neglects the quadratic term $\boldsymbol{c}^2$, which represents the line losses in the distribution network. This simplification is based on the findings in [37], where it is argued that in most practical scenarios, especially in distribution networks, the line losses can be considered relatively small compared to the other terms in the power flow equations. Thus, the quadratic term $\boldsymbol{c}^2$ in (21) is neglected, turning the expression linear in $\boldsymbol{v}_m^2$. This linear expression can further be used to derive a direct relationship between the action vector $\boldsymbol{a}$, which corresponds to the dispatch decision of the aggregators, i.e., $\boldsymbol{a}=\left[p_{1,t}^B,p_{2,t}^B,\ldots,p_{m,t}^B,\ldots,p_{|\mathcal{N}|,t}^B\right]$, and $\boldsymbol{v}_m^2$, which is expressed as:

$$\boldsymbol{M}\boldsymbol{v}^2=\boldsymbol{M}v_0^2\mathbf{1}_{|\mathcal{L}|}+2\left[\boldsymbol{D}\left(\boldsymbol{r}_{mn}\right)\boldsymbol{B}\boldsymbol{T}\left(\boldsymbol{p}_m^{\mathrm{N}}-\boldsymbol{a}\right)+\boldsymbol{D}\left(\boldsymbol{x}_{mn}\right)\boldsymbol{B}\boldsymbol{T}\boldsymbol{q}_m^{\mathrm{N}}\right] \quad (24)$$

C. Safety Layer Formulation

The relationship expressed in (21) is utilized to establish a linear mathematical programming formulation to project potentially unsafe actions, defined by the RL algorithm, into a secure operational region. The primary objective of this formulation is to find the nearest safe action $\hat{\boldsymbol{a}}$ that minimizes the Euclidean distance from the original potentially unsafe action $\boldsymbol{a}$. Thereby, the projection can ensure minimal deviation from the intended control strategy while strictly adhering to operational and safety constraints. The safe action projection is achieved by solving the optimization problem:

$$\hat{\boldsymbol{a}}=\arg\min_{\hat{\boldsymbol{a}}}\left\{\frac{1}{2}\left(\hat{\boldsymbol{a}}-\boldsymbol{a}\right)^2\right\} \quad (25)$$

s.t.

$$\boldsymbol{v}_m^2\mathbf{1}_{|\mathcal{L}|}+2\boldsymbol{M}^{-1}\left[\boldsymbol{D}\left(\boldsymbol{r}_{mn}\right)\boldsymbol{B}\boldsymbol{T}\left(\boldsymbol{p}_m^{\mathrm{N}}-\hat{\boldsymbol{a}}\right)+\boldsymbol{D}\left(\boldsymbol{x}_{mn}\right)\boldsymbol{B}\boldsymbol{T}\boldsymbol{q}_m^{\mathrm{N}}\right]\le\bar{v}^2-\epsilon \quad (26)$$

$$\boldsymbol{v}_m^2\mathbf{1}_{|\mathcal{L}|}+2\boldsymbol{M}^{-1}\left[\boldsymbol{D}\left(\boldsymbol{r}_{mn}\right)\boldsymbol{B}\boldsymbol{T}\left(\boldsymbol{p}_m^{\mathrm{N}}-\hat{\boldsymbol{a}}\right)+\boldsymbol{D}\left(\boldsymbol{x}_{mn}\right)\boldsymbol{B}\boldsymbol{T}\boldsymbol{q}_m^{\mathrm{N}}\right]\ge\underline{v}^2+\epsilon \quad (27)$$

The slack parameter ϵ is introduced to manage the relaxation conditions for the voltage magnitude limits, which compensates for the inaccuracies introduced by the linear model approximation of real voltage magnitudes. By incorporating ϵ, we allow for a buffer in the operational constraints that accommodates potential deviations between the predicted and actual voltage magnitudes. This ensures that the projected actions remain within safe operational boundaries, even when the linear relationship underestimates or overestimates the effects of control actions on the voltage levels.

D. Framework of Proposed DF-SRL Algorithm

The proposed safety layer can project action a_t to safe domains $\hat{a}_t$ during the training and online execution process. The proposed DF-SRL algorithm will update the actor and critic networks based on the collected safe trajectories $\left(s_t,\hat{a}_t,r_t,s_{t+1}\right)$ in the replay buffer R. Therefore, the proposed DF-SRL algorithm redefines the actor-network and critic-network iteration rules by (28) and (29), respectively.

$$\omega\leftarrow\omega+\nabla_\omega\frac{1}{|B|}\sum_{s_t\in B}\left(\min_{i=1,2}\left\{Q_{\theta_i}\left(s_t,\hat{a}_t\right)\right\}\right) \quad (28)$$

$$\min_\theta\sum_{s\in B}\left(r_t+\gamma\min_{i=1,2}\left\{Q_{\theta_i^{\text{target}}}\left(s_{t+1},\hat{a}_t\right)\right\}-Q_{\theta_i}\left(s_t,\hat{a}_t\right)\right)^2 \quad (29)$$

Note that the proposed DF-SRL algorithm for integrating the safety layer is specifically designed to be compatible with off-policy model-free algorithms. The off-policy nature of the proposed DF-SRL algorithm allows it to learn from experiences generated by a behavior policy that differs from the target policy trying to learn. This characteristic is crucial for the integration of the safety layer, as it allows the algorithm to handle the mismatched distribution between the original actions a_t and the safe actions $\hat{a}_t$ without impairing the update performance. Consequently, the safety layer can project potentially unsafe actions into a safe domain, ensuring operational feasibility while maintaining the integrity of the learning process. The proposed DF-SRL algorithm maintains its model-free nature by not explicitly learning the state transition function of the constructed MDP [38].

In addition to the integration of the safety layer, the proposed DF-SRL algorithm introduces significant novelty in the policy iteration and interaction process. More than just filtering actions, the safety layer actively changes the nature of the interaction data that are fed back into the learning process of the RL agent. By modifying the actions before they are executed (and thus the resulting state transitions and rewards), the safety layer ensures that the data used for training are not only rich in terms of learning opportunities but also aligned with operational safety requirements. This leads to an improvement in both the performance and safety of the learned policy.

Algorithm 1 presents the step-by-step procedure of the proposed DF-SRL algorithm, while Fig. 2 illustrates the architecture of proposed DF-SRL algorithm displaying the interaction of the actor and critic models with the environment during the training process. The proposed DF-SRL algorithm composes of actor and critic models and interacts with the environment through the formulated safe layer to ensure the safety and feasibility during the training, as shown in Fig. 2.

Algorithm 1: proposed DF-SRL algorithm

Define the maximum training epoch T and epoch length L
Initialize parameters of functions Q_{θ_i}, $Q_{\theta_i^{\text{target}}}$, and π_ω, and reply buffer R
Define the parameters of the safety layer: $\boldsymbol{D}\left(\boldsymbol{r}_{mn}\right),\boldsymbol{D}\left(\boldsymbol{x}_{mn}\right),\boldsymbol{B},\boldsymbol{T}$, and $\boldsymbol{M}$
for $t=1$ to T **do**
  Sample an initial state s_0 from the initial distribution
  for $l=1$ to L **do**
    Sample an action with exploration noise $a_t\sim\pi_\omega(s_t)+\epsilon$, $\epsilon\sim N(0,\sigma_1)$
    if $\underline{v}<v<\bar{v}$ is not statisfied, **then**
      Project a_t to safe action $\hat{a}_t$ by solving {(25), s.t. (26), (27)}.
    else $\hat{a}_t=a_t$
    Interact with the distribution network and observe the reward r_t and the new state s_{t+1}
    Store the transition tuple $\left(s_t,\hat{a}_t,r_t,s_{t+1}\right)$ in R
    Sample a random mini-batch of B transitions $\left(s_t,\hat{a}_t,r_t,s_{t+1}\right)$ from R
    Update the Q-function parameters by using (29)
    Update the execution policy function parameters by using (28)
    Update the target Q-function parameters using $\theta^{\text{target}}\leftarrow\tau\theta+(1-\tau)\theta^{\text{target}}$

The training process begins by randomly initializing the parameters of the DNN functions Q_θ and $Q_{\theta^{\text{target}}}$, as well as defining the parameters of the safety layer, i. e., $\boldsymbol{D}\left(\boldsymbol{r}_{mn}\right)$, $\boldsymbol{D}\left(\boldsymbol{x}_{mn}\right)$, $\boldsymbol{B}$, $\boldsymbol{T}$, and $\boldsymbol{M}$. For each training epoch, at each time step t, the policy π_ω receives the state s_t and samples an action a_t. The safety layer then assesses whether the action a_t falls within the safe domain. The projection model is activated to project actions to a safe action, denoted as $\hat{a}_t$, only if action a_t could lead to voltage magnitude violations. Next, a transition tuple $\left(s_t,\hat{a}_t,r_t,s_{t+1}\right)$ is compiled and stored in a replay buffer R. A subset B of these samples is subsequently selected and used to update the parameters of the functions Q_{θ_i}, $Q_{\theta_i^{\text{target}}}$, and π_ω, as detailed in Algorithm 1. This iterative procedure continues until the maximum number of epochs is reached, ensuring that the RL agent can efficiently explore the action space without breaching voltage magnitude limits, thereby ensuring operational feasibility.

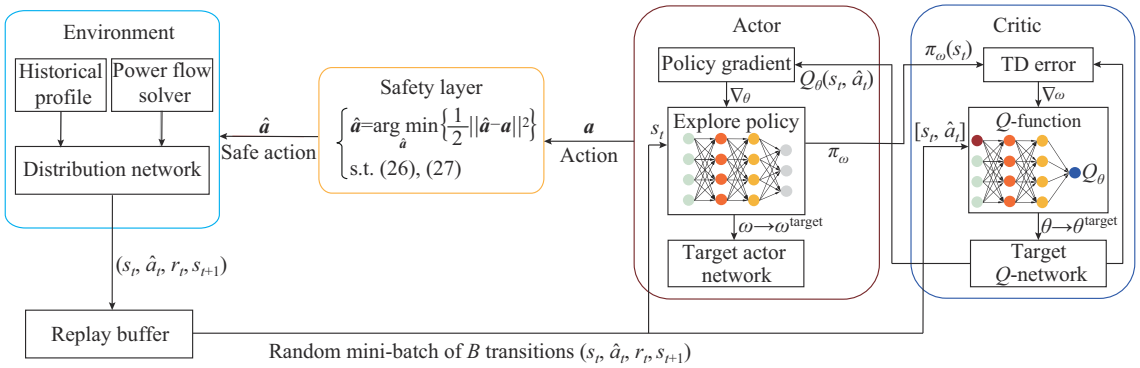

Fig. 2. Architecture of proposed DF-SRL algorithm displaying interaction among actor network, critic network, and safety layer.

IV. Simulation Results and Discussions

A. Simulations Setup, Data, and Implementation

1) Data and Distribution Network Case

To validate the effectiveness of the proposed DF-SRL algorithm, we construct an environment based on a CIGRE residential low-voltage network, as shown in Fig. 3. In this network, each node is associated with an aggregator, and the DSO interacts with them to regulate voltage magnitude based on the availability of flexibility at each node. The training data of PVs, plug-in EVs, and typical residential load are from [3], with a 15-min resolution. The voltage magnitude limits are set to be $\bar{v}=1.05$ p.u. and $\underline{v}=0.95$ p.u.. For the present case study, we assume that the maximal flexibility provided by the aggregator is 50 kW during the operation [3].

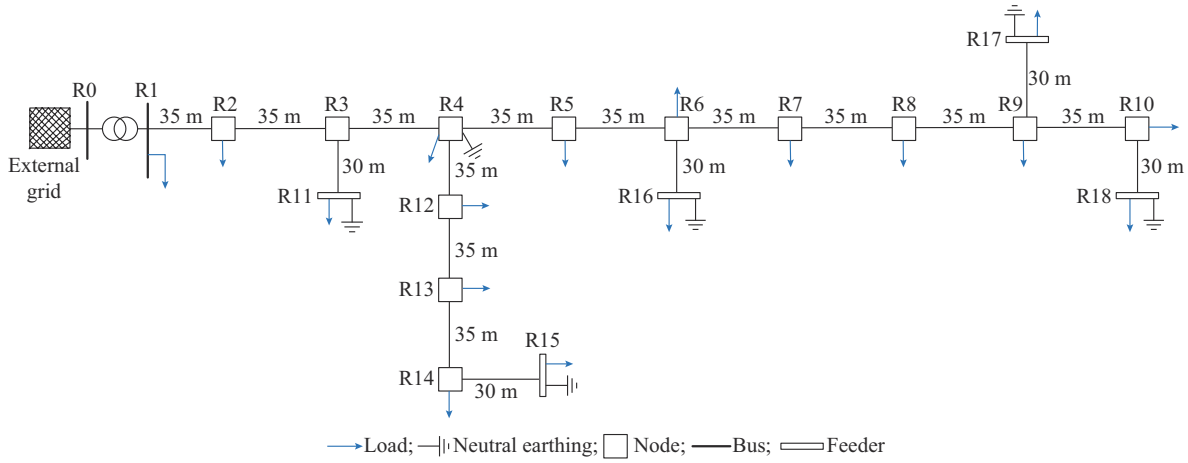

Fig. 3. Modified CIGRE residential low-voltage network.

2) Benchmark Algorithms

To evaluate the performance of the proposed DF-SRL algorithm, we conduct a comparative analysis with several DRL benchmark algorithms, including the state-of-the-art DRL algorithms: DDPG, proximal policy optimization (PPO), TD3, and SAC, as well as a centralized model-based algorithm, i.e., an NLP formulation [3]. The parameters for different DRL algorithms, aggregators, and environment are summarized in Table I. TD3, DDPG, and safe DDPG algorithms are trained with the same hyperparameters as the proposed DF-SRL algorithm. Specifically, linear safety layer training for safe DDPG follows the default implementation in [28]. All implemented algorithms and their (hyper)parameters are available online. Note that while all the DRL benchmark algorithms can make decisions only using current information and achieve online operation, the solution obtained by the NLP formulation requires complete information of the foreseen control period. To train and assess the performance of the DRL benchmark algorithms, we employ validation metrics based on the negative value of total used active power as denoted in (11), and the voltage magnitude violation penalty as specified in (13). These metrics effectively gauge the operational efficiency and constraint adherence of each algorithm.

B. Performance on Training Set

Figure 4 presents a comparative analysis of the average total reward as in (17), the summation of negative value of total used active power (or the first term of reward in (17)), and the cumulative penalty for voltage magnitude violations (or the second term of reward in (17)) during the training process for the proposed DF-SRL and the DRL benchmark algorithms. Results shown in Fig. 4 for each algorithm are obtained as an average of over five executions. The average total reward increases rapidly during the training, while voltage magnitude violations decrease significantly at the beginning.

TABLE I
PARAMETERS FOR DRL ALGORITHMS, AGGREGATORS, AND ENVIRONMENT

Item	Parameter
DF-SRL	$\gamma = 0.995$
	Optimizer adopts Adam
	Learning rate is 6×10^{-4}
	Batch size is 512, replay buffer is 4×10^5
SAC	$\gamma = 0.995$
	Optimizer adopts Adam
	Learning rate is 6×10^{-4}
	Batch size is 512, replay buffer is 4×10^5
	Entropy is fixed
PPO	$\gamma = 0.995$
	Optimizer adopts Adam
	Learning rate is 6×10^{-4}
	Batch size is 4096
Aggregator	$\bar{p}^B = 50$ kW, $\underline{p}^B = -50$ kW
Environment	Reward $\sigma = 400$
	Voltage limit $\bar{v} = 1.05$ p.u., $\underline{v} = 0.95$ p.u.

As depicted in Fig. 4(b), it is noteworthy that the negative values of total used active power for DDPG and TD3 algorithms (with soft penalty) eventually converges around −1.7 MW, while that of SAC and safe DDPG converges around −2.4 and −4.3 MW, respectively. Compared with these DRL benchmark algorithms, the negative values of total used active power of the proposed DF-SRL algorithm is significantly lower, at approximately −0.5 MW. Figure 4(c) reveals another stark contrast in the cumulative penalty for voltage magnitude violations. Throughout the training process, the proposed DF-SRL algorithm consistently enforces the constraints without any voltage magnitude violations, whereas the DRL benchmark algorithms experience failures in satisfying the constraints after reaching convergence at 1000 episodes. This disparity can be attributed to the safety layer in the proposed DF-SRL algorithm, which adjusts unsafe actions during training. In comparison, the DRL benchmark algorithms initially grapple with low-quality actions due to the random initialization of the DNN parameters, leading to many initial violations. Then, based on the guidance of the penalty term of the reward function, the DDPG, TD3, and PPO algorithms reduce the voltage magnitude violations to a small value after about 200 episodes. Conversely, the SAC algorithm exhibits slower training efficiency, achieving smaller violation values only at the end of training (1000 episodes). This behavior can be attributed to the complex exploration policy used by the SAC algorithm. The safe DDPG algorithm maintains relatively smaller violation values at the beginning compared with other algorithms (e.g., TD3) with a soft penalty. Nevertheless, it fails to enforce violations caused by the poor quality of the safety layer, trained based on the data collected from random policy-environment interaction. Moreover, the infeasible safety layer also leads the action project in the wrong direction, impacting the data quality in the replay buffer, which causes a worse performance compared with the standard counterpart (i.e., DDPG), as shown in Fig. 4(b).

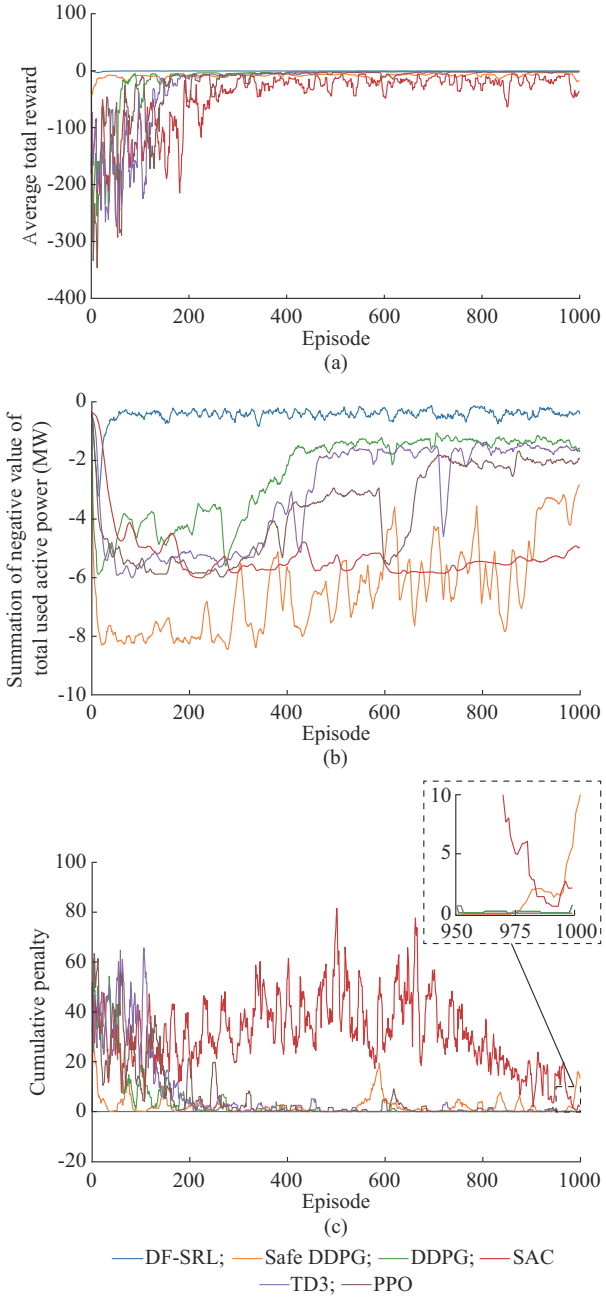

Fig. 4. Comparative analysis results of different algorithms. (a) Average total reward. (b) Summation of negative value of total used active power. (c) Cumulative penalty for voltage magnitude violations.

C. Performance and Constraint Enforcement Capabilities on Testing Set

Figure 5 displays the results of voltage magnitude of different nodes before and after the regulation of the proposed DF-SRL, safe DDPG, and TD3 algorithms, and NLP formulation during a typical day in the test dataset. As TD3 algorithm performs best among all DRL algorithms, we use the TD3 algorithm as a benchmark. In the specific scenario of nodes 11, 16, 17, and 18 of the network operating under severe undervoltage during afternoon and night, the proposed DF-SRL algorithm effectively maintains the voltage magnitude within the technical limits throughout the entire opera-

tion period. Notably, the safe DDPG algorithm fails to maintain the voltage magnitude within the technical limits between 20:00-21:00. This is due to the inherent limitations of the trained linear safety layer, which performs poorly in the distribution network environment with complex dynamics and multiple constraints involved. Similarly, DRL benchmark algorithms, for instance, TD3 algorithm, trained with a soft penalty, cannot provide certified feasibility after convergence. Furthermore, the operational cost associated with the regulation of the proposed DF-SRL algorithm is 0.76 MW, a significant reduction of 17.7% compared with that of TD3 and safe DDPG algorithms. This reduction can be attributed to the high-quality training data provided by the expert-knowledge-based safety layer in the proposed DF-SRL algorithm. Compared with the optimal solution obtained by solving the NLP formulation with a perfect forecast, the proposed DF-SRL algorithm demonstrates a modest error rate of 10.6%.

Table II presents the average total error in operational cost, the average number of voltage magnitude violations (including over- and under-voltage violations), and the average total computational time for the proposed DF-SRL and DRL benchmark algorithms assessed over 30 unseen test days. As illustrated in Table II, the proposed DF-SRL algorithm consistently upholds voltage magnitude constraints while achieving a marked reduction in average error relative to the solution obtained by the NLP formulation with perfect forecast. In general, the proposed DF-SRL algorithm performs the best among all the algorithms with the lowest average error of 11.6%. In contrast, the TD3 algorithm underperforms with an error rate of 35.9%, violating voltage magnitude constraints around 14 time steps. Other DRL algorithms such as the DDPG, PPO, and SAC algorithms register higher errors at 37.2%, 44.3%, and 56.1%, respectively. With a trained linear safety layer, the safe DDPG algorithm fails to enforce voltage magnitude constraints while performing worse than the standard DDPG algorithm. This is because the trained safety layer in the safe DDPG algorithm cannot accurately track the relationship between state, action, and multiple constraints. As anticipated, due to the computation of the safety layer, the proposed DF-SRL algorithm requires more computational resources compared with other DRL algorithms. Despite this, the proposed DF-SRL algorithm remains a viable option for real-time operation as it takes less than 29 s for one day (96 time steps) execution.

D. Sensitive Analysis

The proposed DF-SRL algorithm capitalizes on the linear relationship between the voltage magnitude and the actions. Nevertheless, the power flow formulation can introduce errors due to the approximation assumptions. The safety layer formulation introduces the slack parameter ϵ to overcome this. Primarily, ϵ should be determined by the upper error boundary for the DistFlow model compared with the actual voltage magnitude. As the final value used for ϵ influences the feasibility and optimality of the actions defined by the proposed DF-SRL algorithm, this subsection presents an in-depth sensitivity analysis of the slack parameter ϵ.

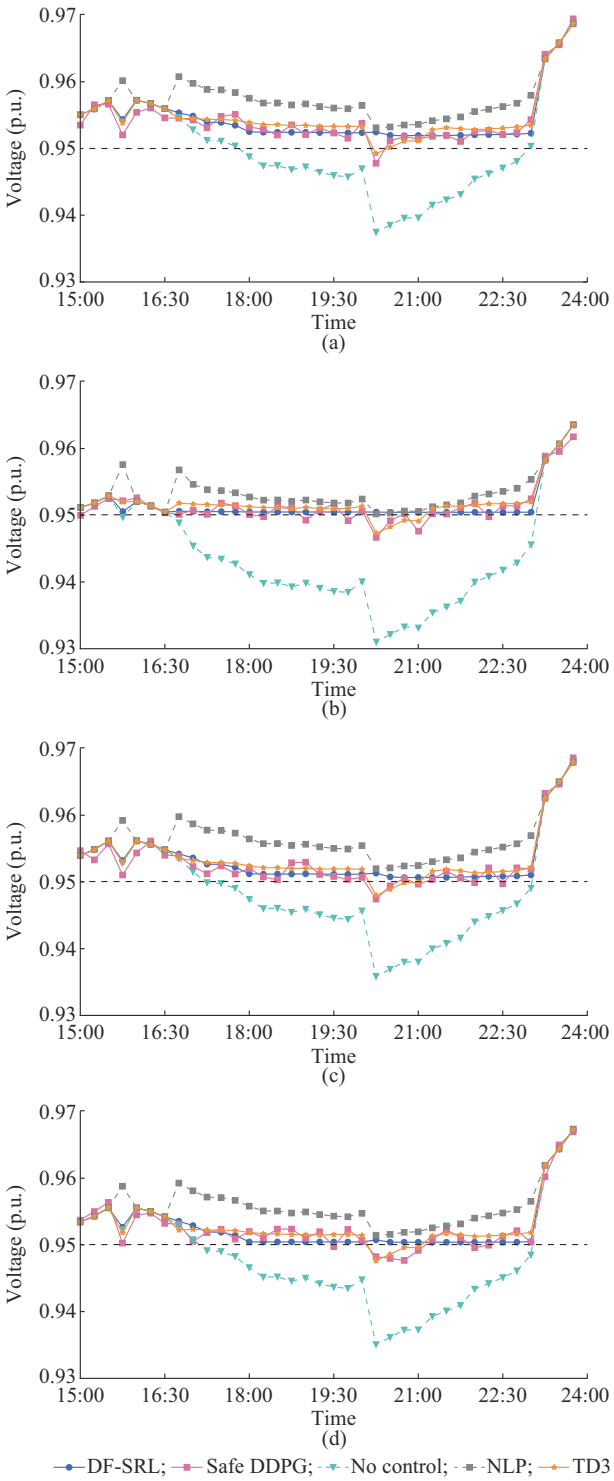

Fig. 5. Voltage magnitude of different nodes before and after regulation of DF-SRL, safe DDPG, and TD3 algorithms, and NLP formulation. (a) Node 11. (b) Node 15. (c) Node 17. (d) Node 18.

Figure 6 illustrates the convergence performance of the proposed DF-SRL algorithm for different values of the slack parameter ϵ. At $\epsilon = 0.001$, the performance of the proposed DF-SRL algorithm is markedly diminished after convergence. In this case, the total active power provided by the aggregators is relatively low compared with the cases when ϵ takes the value of 0.002 or 0.005.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT DRL ALGORITHMS

Algorithm	Average total error (%)	Average number of voltage magnitude violations	Average total computational time (s)
DF-SRL	11.6 ± 0.0	0	29.0 ± 2.4
Safe DDPG	67.1 ± 5.5	19 ± 2	25.0 ± 0.7
DDPG	37.2 ± 1.2	15 ± 4	15.7 ± 0.2
TD3	35.9 ± 1.5	14 ± 4	15.7 ± 0.2
SAC	56.1 ± 3.4	23 ± 4	16.0 ± 0.1
PPO	44.3 ± 1.1	12 ± 1	15.4 ± 0.6

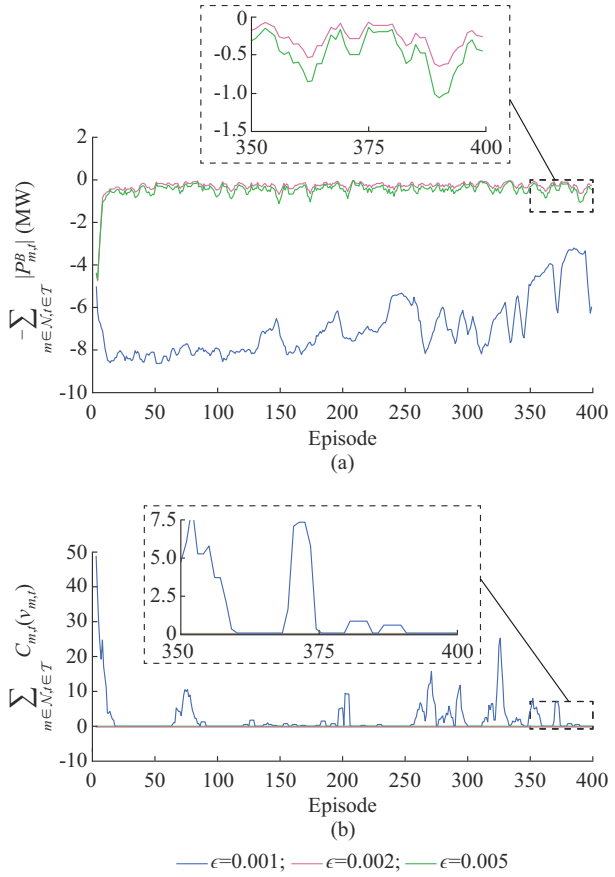

Fig. 6. Convergence performance of proposed DF-SRL algorithm for different ϵ. (a) Total flexible active power. (b) Number of voltage magnitude violations.

Additionally, in the case of $\epsilon = 0.001$, the proposed DF-SRL algorithm fails to ensure the feasibility of the decided solutions during training, whereas in the cases with ϵ set at 0.002 or 0.005, all operational constraints can be successfully enforced. In general, a low value of ϵ can make the safe solution of the linear projection model infeasible. Consequently, the resolved safe solution may cause voltage magnitude violations during training, leading to sub-optimal performance after projection. If the proposed DF-SRL algorithm is executed with ϵ being 0.002 or 0.005, significant performance improvements in optimality and feasibility are observed, as illustrated in Fig. 6. Furthermore, the optimality score experiences a modest increase, at around 5%, when ϵ is reduced from 0.005 to 0.002. This can be attributed to the fact that a higher ϵ constrains the solution space in the action projection model, subsequently affecting the solution quality during training. The calibration of the slack parameter ϵ is intrinsically linked to the linear error inherent in the safety layer, which is pivotal for the efficacy of the proposed DF-SRL algorithm. This calibration ensures that the relaxations provided by ϵ comprehensively cover the linearization errors, thus maintaining the integrity of the safety layer across varying operational scenarios. In the following section, we conduct a detailed scalability analysis to further explore the range of errors induced by the linearization process, providing a quantitative foundation to refine the selection of ϵ across different network sizes [37].

V. SCALABILITY ANALYSIS

The scalability of the proposed DF-SRL algorithm is fundamentally determined by the effectiveness of the DistFlow linearization process. This linearization approximation is essential for mapping the actions from the DRL to safe operational domains. Substantial linearization errors can cause inaccuracies within the safety layer, misguiding action projection, compromising policy iterations, and ultimately degrading the overall efficacy of the algorithm.

Figure 7 presents the observed voltage magnitude errors of DistFlow for different network sizes.

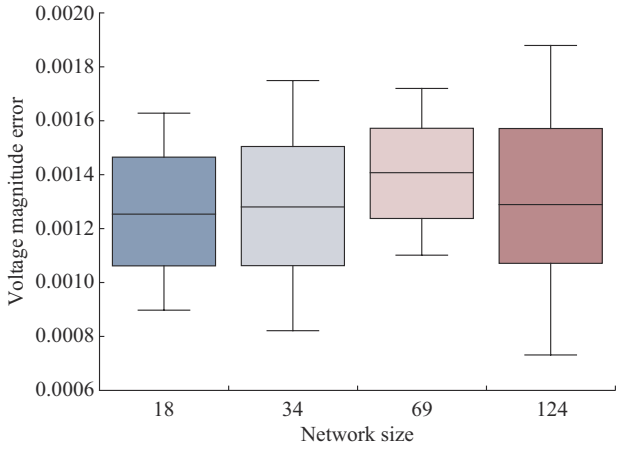

Fig. 7. Voltage magnitude errors of DistFlow on 18-, 34-, 69-, and 124-node distribution networks.

We collect voltage magnitudes from all nodes within distribution networks of 18, 34, 69, and 124 nodes and calculate the deviations between the DistFlow approximations and actual voltage magnitudes in one year's data. The voltage magnitude error in the 18-node distribution network ranges from 0.00089 to 0.00163. In the 34-node distribution network, the error ranges from 0.00082 to 0.00175. The 69-node distribution network experiences an error range of 0.0011 to 0.00172, and the 124-node distribution network experiences an error range from 0.00073 to 0.00188. Although the largest distribution network exhibits a broader range of error, the maximum error does not exceed 0.002, suggesting that setting an error threshold of $\epsilon = 0.002$ effectively accommodates the inaccuracies induced by the linearization across

all tested distribution networks. The results demonstrate the robustness of the DistFlow model, which forms a solid foundation for the safety layer, facilitating its application across diverse distribution network configurations. This generalizability ensures that with precise data on the parameters and topology of the distribution network, the safety layer can be tailored to maintain its accuracy and relevance, regardless of the specific characteristics of the distribution network.

VI. Conclusion

The DF-SRL algorithm developed in this paper demonstrated its superior performance in handling voltage magnitude constraints while maintaining performance efficiency. In the testing phase, the DF-SRL algorithm effectively maintains voltage magnitude constraints even under severe conditions (e.g., under-voltage problem caused by extreme loading at the marginal node of the network), resulting in an operational cost reduction of 17.7% compared with the benchmark algorithms, while ensuring feasibility throughout the entire operation period. Specifically, the DF-SRL algorithm enforced voltage magnitude constraints without violations, even in unseen data. This is attributable to the safety layer embedded in the DF-SRL algorithm, designed to filter out unsafe actions during the training phase, thus eliminating voltage violations. The sensitivity analysis of the slack parameter ϵ found that its value significantly impacts the optimality and feasibility of the DF-SRL algorithm. We found that $\epsilon = 0.002$ provides an optimal balance between rigorously enforcing the constraints and achieving the highest performance score. The scalability analysis conducted across various network sizes demonstrated conclusively that the DF-SRL algorithm maintains high performance and accuracy in voltage magnitude regulation, effectively substantiating its utility and robustness for practical, large-scale applications. Its versatility allows for integration with any off-policy DRL algorithm, facilitating the resolution of continuous control challenges within distribution network operations underpinned by state-wise constraints.

References

[1] S. J. Davis, N. S. Lewis, M. Shaner *et al*., "Net-zero emissions energy systems," *Science*, vol. 360, p. 9793, Jun. 2018.

[2] A. G. Trojani, M. S. Moghaddam, and J. M. Baigi, "Stochastic security-constrained unit commitment considering electric vehicles, energy storage systems, and flexible loads with renewable energy resources," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 5, pp. 1405-1414, Sept. 2023.

[3] A. Fu, M. Cvetkovic, and P. Palensky, "Distributed cooperation for voltage regulation in future distribution networks," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4483-4493, Nov. 2022.

[4] X. Chen, E. Dall'Anese, C. Zhao *et al*., "Aggregate power flexibility in unbalanced distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 258-269, Jan. 2020.

[5] C. Li, K. Zheng, H. Guo *et al*., "Intra-day optimal power flow considering flexible workload scheduling of IDCs," *Energy Reports*, vol. 9, pp. 1149-1159, Sept. 2023.

[6] Y. Li, Y. Gu, G. He *et al*., "Optimal dispatch of battery energy storage in distribution network considering electrothermal-aging coupling," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3744-3758, Sept. 2023.

[7] M. Glavic, "(Deep) Reinforcement learning for electric power system control and related problems: a short review and perspectives," *Annual Reviews in Control*, vol. 48, pp. 22-35, Oct. 2019.

[8] S. Hou, E. M. Salazar, P. P. Vergara *et al*., "Performance comparison of deep RL algorithms for energy systems optimal scheduling," in *Proceedings of 2022 IEEE PES Innovative Smart Grid Technologies Conference Europe*, Novi Sad, Serbia, Oct. 2022, pp. 1-6.

[9] M. Xia, F. Chen, Q. Chen *et al*., "Optimal scheduling of residential heating, ventilation and air conditioning based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 5, pp. 1596-1605, Sept. 2023.

[10] P. P. Vergara, M. Salazar, J. S. Giraldo *et al*., "Optimal dispatch of PV inverters in unbalanced distribution systems using reinforcement learning," *International Journal of Electrical Power and Energy Systems*, vol. 136, p. 107628, Mar. 2022.

[11] S. Wang, J. Duan, D. Shi *et al*., "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4644-4654, Nov. 2020.

[12] H. Ding, Y. Xu, B. C. S. Hao *et al*., "A safe reinforcement learning approach for multi-energy management of smart home," *Electric Power Systems Research*, vol. 210, p. 108120, Sept. 2022.

[13] E. M. S. Duque, J. S. Giraldo, P. P. Vergara *et al*., "Community energy storage operation via reinforcement learning with eligibility traces," *Electric Power Systems Research*, vol. 212, p. 108515, Nov. 2022.

[14] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, pp. 1437-1480, Jul. 2015.

[15] S. Zhang, R. Jia, H. Pan *et al*., "A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid," *Applied Energy*, vol. 348, p. 121490, Oct. 2023.

[16] H. Ding, Y. Xu, B. C. S. Hao *et al*., "A safe reinforcement learning approach for multi-energy management of smart home," *Electric Power Systems Research*, vol. 210, p. 108120, Sept. 2022.

[17] X. Yang, H. He, Z. Wei *et al*., "Enabling safety-enhanced fast charging of electric vehicles via soft actor critic-Lagrange DRL algorithm in a cyber-physical system," *Applied Energy*, vol. 329, p. 120272, Jan. 2023.

[18] H. Cui, Y. Ye, J. Hu *et al*., "Online preventive control for transmission overload relief using safe reinforcement learning with enhanced spatial-temporal awareness," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 517-532, Jan. 2024.

[19] J. Achiam, D. Held, A. Tamar *et al.*, "Constrained policy optimization," in *Proceedings of International Conference on Machine Learning*, Sydney, Australia, Aug. 2017, pp. 22-31.

[20] H. Li and H. He, "Learning to operate distribution networks with safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 1860-1872, May 2022.

[21] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2427-2439, May 2020.

[22] W. Zhao, T. He, R. Chen *et al.* (2023, Feb.). State-wise safe reinforcement learning: a survey. [Online]. Available: https://www.ijcai.org/proceedings/2023/763

[23] S. Hou, E. M. S. Duque, P. Palensky *et al.* (2023, Jul.). A constraint enforcement deep reinforcement learning framework for optimal energy storage systems dispatch. [Online]. Available: https://arxiv.org/abs/2307.14304

[24] S. Hou, P. P. Vergara, E. M. S. Duque *et al*., "Optimal energy system scheduling using a constraint-aware reinforcement learning algorithm," *International Journal of Electrical Power & Energy Systems*, vol. 152, p. 109230, Oct. 2023.

[25] W. Cui, J. Li, and B. Zhang, "Decentralized safe reinforcement learning for inverter-based voltage control," *Electric Power Systems Research*, vol. 211, p. 108609, Oct. 2022.

[26] W. Wang, N. Yu, Y. Gao *et al*., "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008-3018, Jul. 2020.

[27] M. Zhang, G. Guo, T. Zhao *et al*., "DNN assisted projection based deep reinforcement learning for safe control of distribution grids," *IEEE Transactions on Power Systems*, vol. 39, no. 4, pp. 5687-5698, Jul. 2024.

[28] G. Dalal, K. Dvijotham, M. Vecerik *et al.* (2018, Jan.). Safe exploration in continuous action spaces. [Online]. Available: https://arxiv.org/pdf/1801.08757v1

[29] M. Eichelbeck, H. Markgraf, and M. Althoff, "Contingency-constrained economic dispatch with safe reinforcement learning," in *Proceedings of 2022 21st IEEE International Conference on Machine Learning and Applications*, Nassau, Bahamas, Dec. 2022, pp. 597-602.

[30] P. Kou, D. Liang, C. Wang *et al*., "Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks," *Applied Energy*, vol. 264, p. 114772, Apr. 2020.
[31] T. H. Pham, G. de Magistris, and R. Tachibana, "OptLayer – practical constrained optimization for deep reinforcement learning in the real world," in *Proceeding of 2018 IEEE International Conference on Robotics and Automation*, Brisbane, Australia, May 2018, pp. 6236-6243.
[32] E. D. Klenske and P. Hennig, "Dual control for approximate bayesian reinforcement learning," *Journal of Machine Learning Research*, vol. 17, pp. 1-30, Aug. 2016.
[33] X. Zhang, T. Yu, Z. Pan *et al*., "Lifelong learning for complementary generation control of interconnected power grids with high-penetration renewables and EVs," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4097-4110, Jul. 2018.
[34] V. Mnih, K. Kavukcuoglu, D. Silver *et al*., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, Feb. 2015.
[35] T. Lilicrap, J. Hunt, A. Pritzel *et al.*, "Continuous control with deep reinforcement learning," in *Proceedings of International Conference on Learning Representations*, San Juan, Puerto Roco, May 2016, pp. 1221-234.
[36] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of International Conference on Machine Learning*, Stockholm, Sweden, Jul. 2018, pp. 1587-596.
[37] E. Schweitzer, S. Saha, A. Scaglione *et al*., "Lossy DistFlow formulation for single and multiphase radial feeders," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1758-1768, May 2020.
[38] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, p. 1054, Sept. 1998.

Shengren Hou received the B.S. degree in electric engineering from Northeast Electric Power University, Jilin, China, in 2018, and the M.S. degree in electric engineering from Guangxi University, Guangxi, China, in 2021. He is currently pursuing the Ph.D. degree at the Delft University of Technology, Delft, The Netherlands. His main research interests include active distribution network optimization control, short-term electricity market arbitrage, and reinforcement learning.

Aihui Fu received the B.S. degree in electric engineering from China Agricultural University, Beijing, China, 2015, and the M.S. degree in electrical engineering from Shandong University, Jinan, China, in 2018. She is currently pursuing the Ph. D. degree at the Delft University of Technology Delft, The Netherlands. Her research interests include distributed optimization control, power system stability analysis, renewable energy resource, and battery energy storage.

Edgar Mauricio Salazar Duque received the B.E. degree in electrical and electronic engineering from the Universidad de Los Andes, Bogotá, Colombia, in 2008, the M.Sc. degree (cum laude) in smart electrical grids and systems from the Kungliga Tekniska Högskolan (KTH), Stockholm, Sweden, and the Technical University of Eindhoven, Eindhoven, The Netherlands, in 2018. He is currently working towards a Ph.D. degree in the electrical energy systems group at the Technical University of Eindhoven. His research interests include data analysis, and application of machine learning techniques on power distribution grid for planning and operation.

Peter Palensky received the M.Sc., Ph.D., and Habilitation degrees from Vienna University of Technology, Vienna, Austria, in 1997, 2001, and 2015, respectively. He is currently a Full Professor of intelligent electric power grids and the Head of the Electrical Sustainable Energy Department, Delft University of Technology, Delft, The Netherlands. His research interests include energy automation network, smart grid, and modeling of intelligent energy system.

Qixin Chen received the Ph. D. degree from the Department of Electrical Engineering, Tsinghua University, Beijing, China, in 2010, where he is currently a Tenured Professor. His research interests include electricity market, power system economics and optimization, low-carbon electricity, and data analytics in power system.

Pedro P. Vergara received the B.Sc. degree (with honors) in electronic engineering from the Universidad Industrial de Santander, Bucaramanga, Colombia, in 2012, and the M. Sc. degree in electrical engineering from the University of Campinas, UNICAMP, Campinas, Brazil, in 2015. In 2019, he received his Ph. D. degree from the University of Campinas, UNICAMP, and the University of Southern Denmark, SDU, Denmark, funded by the Sao Paulo Research Foundation (FAPESP). In 2019, he joined the Eindhoven University of Technology, TU/e, Eindhoven, The Netherlands, as a Postdoctoral Researcher. In 2020, he was appointed as Assistant Professor at the Intelligent Electrical Power Grids (IEPG) Group at Delft University of Technology, Delft, The Netherlands. He received the Best Presentation Award at the Summer Optimization School in 2018 organized by the Technical University of Denmark (DTU), Copenhagen, Denmark, and the Best Paper Award at the 3rd IEEE International Conference on Smart Energy Systems and Technologies (SEST) in Turkey in 2020. His main research interests include the development of algorithms for control, planning, and operation of electrical distribution systems with high penetration of low-carbon energy resources (e.g, electric vehicle, PV system, electric heat pump) using optimization and machine learning approaches.