Data-Driven Offline-Online Voltage Regulation for Distributed PV Integration

Wangqing Mao[a,b], Hongxing Ye[a,*], Shaojie Zhang[a], Hangyu Zhao[b], Yinyin Ge[a]

[a] *Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China*
[b] *State Grid Suqian Power Supply Company, Suqian, 223800, Jiangsu, China*

Abstract

The distributed photovoltaics (PV) have experienced rapid development in recent decades, resulting voltage-violation challenge in distribution systems with high-level PV. Based on devices' control timescales, this work presents a novel offline-online voltage optimization framework. In the offline stage, it determines monthly-changing positions for transformer taps and decentralized control policies for capacitor banks. In online stage, PV reactive power and SVG are optimally adjusted, and capacitor control is triggered in real time based on predetermined control policy. We propose a practically-feasible data-driven power flow model incorporating transformer tap positions. Random forest is employed to learn the decentralized control policy. The case studies are performed with a real-world distribution system. The results indicate the proposed technique is scalable to large-scale systems, and the approach decreases PV curtailment by up to 20.60% in a real-world distribution system.

Keywords: Voltage Violation, Distributed PV, Data Driven, Offline-Online Optimization, Decentralized Control policy

1. Introduction

In recent years, photovoltaics (PV) has developed rapidly globally. In particular, the large-scale integration of distributed photovoltaic has steadily increased its proportion in distribution systems. According to the International Renewable Energy Agency(IRENA), the global newly installed PV capacity reached 452 GW in 2024 IRENA (2025). Among them, China contributed 278 GW,

[*] Corresponding author: yehxing@xjtu.edu.cn (Hongxing Ye)

including 118 GW from distributed PV installations Administration (2025). In 2024, the United States added approximately 38 GW of PV capacity, while the 27 countries of the European Union added about 58 GW of PV capacity IRENA (2025). By 2030, IRENA expects global total installed renewable power generation capacity reaches 11,174 GW in the 1.5°C Scenario. Specifically, installed solar PV capacity is expected to rise to more than 5,400 GW Agency (2023).

The rapid growth and high penetration of distributed PV have posed significant challenges to distribution system operations Correa and Vieira (2024). According to Kirchhoff's Voltage Law (KVL), the nodal voltage must rise for the PV power to be injected into the distribution networks. Therefore, high-penetration PV distribution systems often confront over-voltage issues Nazih et al. (2025). In the meantime, the inherent intermittency and volatility of the PV output can also cause significant voltage deviations. The voltage control capability is thus fundamentally important for the system with high-level PV. On 28 April 2025, Spain and Portugal were hit by the most serious blackout in the European power grid in more than two decades Panel (2025). Enhancing voltage control and protection against oscillations and using power electronics for voltage management are recommended by the Committee for analysis of 4-28 Electricity Crisis the April 28 Electricity Crisis (2025).

In recent years, significant research has focused on voltage regulation and optimization in distribution systems. The pioneering work Baran and Wu (1989) introduces an optimization approach to voltage support and develops a DistFlow model widely adopted in the literature. Authors in Xu et al. (2017) present a two-stage Voltage/Var optimization method, in which control capacitor bank and transformer are scheduled in the hourly timescale and inverters are adjusted in the minutes timescale. Reference Hu et al. (2022) proposes to minimize network losses, based on a mixed integer second-order cone program. By leveraging flexible resources, the voltage control capability is enhanced through robust optimization, thereby increasing hosting capacity Zhang et al. (2023). Reference Meng et al. (2024) presents a distributed control scheme for PV inverters considering uncertainty. The authors in Wang et al. (2025) develop a double-layer control model to regulate reactive power and voltage. The upper level coordinates the network, while the lower level focuses on local reactive power compensation. Voltage violations are mitigated with a ramping-based variable timescales model in Zhang et al. (2025). To address the computational challenge, researchers relax power flow equations into second-order cone (SOC) constraints Jabr (2006); Chowdhury et al. (2025). However, it still exhibits slow computation times when solving large-scale problems Byeon et al. (2024). In particular, with the introduction of discrete variables, such as actions for tap-changing transformers and capacitor banks, mixed-integer second-order cone programming (MISCOP) experiences even worse computational performance.

With the rapid development of artificial intelligence, sophisticated data-driven methods have emerged to address the voltage regulation problem in recent years. Among them, the authors in Duan et al. (2020) employ deep reinforcement learning to develop voltage control strategies using model-free agents. This approach relies on real-time measurement data to make control decisions. Clustering and long short-term memory neural networks are utilized in Wang et al. (2021) to develop a robust equivalent model of the distribution system. Short-term voltage stability is assessed by deep learning approaches without time-domain simulations in Huang et al. (2021); Li et al. (2024). In Hong and Zhang (2022), recursive kernel regression and interior point methods are integrated to optimize voltage online. In Zhang et al. (2024), a deep deterministic policy gradient (DDPG)-based approach is proposed to regulate distribution voltage with distributed resources. The Markov decision process and quadratic programming are utilized to determine optimal or near-optimal solutions. Authors in Priyadarshi et al. (2026) present a recurrent neural network based long short-term memory (LSTM) approach to consider voltage stability.

However, practical voltage control in distribution systems still faces unresolved challenges. Many works assume that control assets are equipped with mature communication infrastructures. However, a significant number of existing low-voltage transformers and capacitor banks lack communication functionality. Control signals cannot be dispatched to these legacy devices within seconds or hours. On the other hand, deterministic methods suffer from slow computational performance for large-scale distribution networks. Meanwhile, learning-based methods may be criticized for their limited explanatory insights. This work tries to address the challenges mentioned above. The key contributions are summarized as follows:

(1) A novel offline-online voltage optimization framework is proposed to regulate voltage for high penetration distributed PVs. Existing voltage regulations typically account for short-term (hourly) and real-time (seconds) timescales Xu et al. (2017); Zhang et al. (2024). However, many

assets, such as low-voltage transformers and capacitor banks without communication function, are changed on a monthly basis in practice. The proposed model aims to improve voltage quality over months by coordinating assets in monthly and seconds timescales. The offline stage optimizes transformer taps and the control policy for capacitors on a monthly timescale, managing long-term variations. The online stage adjusts reactive power in real time for devices with remote control capabilities, fine-tuning the operating point in response to real-time disturbances.

(2) A novel data-driven approximation and Random Forest are introduced to solve the large-scale offline-online voltage regulation problem. A rising challenge is that the model size significantly increases when coordinating monthly and second timescales. There is an emerging research need for fast and accurate solution approach that considers transformer taps. We introduce a data-driven linear formulation to approximate power flow with transformers. Random Forest is introduced to learn the optimal control policy for capacitor banks. By training the policy using a Random Forest model, the learning output mimics coordination with other devices, even without available communication. The Synthetic Minority Over-sampling Technique (SMOTE) is employed to address the sample imbalance challenge for voltage violations.

The rest of the paper is organized as follows. Section II presents the framework of offline-online voltage regulation. A novel data-driven methods is developed in Section III. Section IV validates the effectiveness of the proposed method through case studies. Finally, Section V concludes the paper.

2. Offline-Online Multi-level Voltage Optimization

2.1. Background

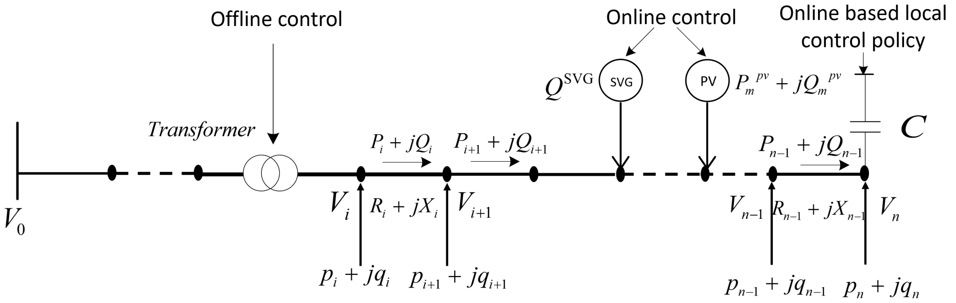


Figure 1: Distribution Network with Voltage Regulation Resources

In the high-penetration PV distribution network, the voltage violation can be observed from the power flow equation. An illustrative distribution network is depicted in Figure 1. Let V_i denote the voltage magnitude for node i, and let lowercase p_i and q_i denote the real and reactive power injection at node i. Let uppercase P_i and Q_i denote the real and reactive power from node i to $i+1$, respectively. Let $R_i + jX_i$ denote the impedance of the line $i-1$ to i. According to the DistFlow power flow equation Baran and Wu (1989); Yeh et al. (2012); Carvalho et al. (2022), the voltage relationship between node n and node $n-1$ can be expressed as:

$$P_{i+1} = P_i - R_i \frac{P_i^2 + Q_i^2}{V_i^2} + p_{i+1} \tag{1}$$

$$Q_{i+1} = Q_i - X_i \frac{P_i^2 + Q_i^2}{V_i^2} + q_{i+1} \tag{2}$$

$$V_{i+1}^2 = V_i^2 - 2(p_i R_i + q_i X_i) + (R_i^2 + X_i^2) \frac{P_i^2 + Q_i^2}{V_i^2} \tag{3}$$

PV generation is included in net power injection p_i. According to Equation (3), the voltage difference between node i and $i+1$, i.e., $V_i^2 - V_{i+1}^2$, rises if the PV generation increases in p_i. Equation (3) also shows that the voltage can be regulated through reactive power adjustment, i.e., $q_i X_i$. Therefore, VAR devices, such as PV inverter and SVG, can be employed to mitigate the voltage violation. In the meantime, we can also directly regulate the voltage via transformers.

2.2. *Offline Stage*

In practice, low-voltage transformer tap changers and the low-voltage capacitor bank control policy can be adjusted every month to three months. However, these offline planning cannot accommodate real-time fluctuations, leading to potential voltage violations during rapid changes. In contrast, many newly installed PV inverters and SVG can be dispatched in real time. Different voltage regulation devices have various characteristics, especially timescales. They are illustrated in Figure 2. The PV inverter and the SVG can be regulated in seconds. When the control policy is given, the local capacitor banks can also respond in real time. In contrast, the low-voltage transformer tap and the capacitor control policy can be set every month to three months in practice.

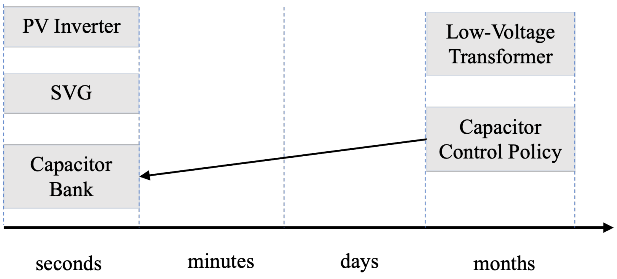


Figure 2: Timescales of devices for voltage regulation.

The offline stage is to determine the optimal transformer positions and the control policy for capacitor banks. The core challenge is that the offline decision must be made in anticipation of its impact on future online decisions. In this subsection, we first present the transformer model and control policy.

The transformer is treated as an equivalent line as shown in Figure 3. Let Z_i be the series impedance of transformer and K_i be the ratio.

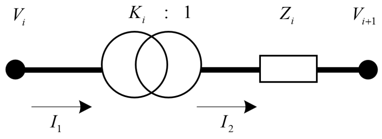


Figure 3: Equivalent line model for the transformer.

The voltage drop can be written as

$$\begin{aligned} V_{i+1}^2 &= |V_i/K_i - K_i I_1 Z_i|^2 \\ &= (V_i/K_i)^2 - 2Re\,(V_i(I_1 Z_i)^*) + |K_i I_1 Z_i|^2 \\ &= \frac{V_i^2}{K_i^2} - 2(P_i R_i + Q_i X_i) + K_i^2 (R_i^2 + X_i^2) \frac{P_i^2 + Q_i^2}{V_i^2} \end{aligned} \tag{4}$$

Low-voltage capacitor banks are widely used in the distribution network. Many of them do not have a communication infrastructure. The low-voltage capacitor action is triggered by local voltage/current measurements following the control policy. Hence, it is attractive to find the optimal or near optimal control policy based only on local information. We write the control policy as follows

$$y_i = G(p_i, q_i, v_i), \quad q_i^{\text{CB}} = y_i q_i^{\text{CB,COM}}$$

where y_i is a discrete variable and G is the control policy, which is a function of p_i, q_i and v_i. q_i^{CB} is the reactive power of the capacitor bank at node i.

2.3. Online Stage

The online stage determines the optimal output of PV inverters and SVGs. It should be noted that the capacitor bank responds in real time, although its control policy is determined offline. The optimal output of PV inverters can be formulated as follows:

$$-\lambda p_i^{\text{pv,actual}} \leq q_i^{\text{pv}} \leq \lambda p_i^{\text{pv,actual}} \tag{5}$$

where q_i^{pv} denotes the reactive power output of PV at node i. Equation (5) means that the reactive power of the PV generator should be within a certain ratio of the active power. Following GB/T 29319-2024 Standardization Administration of China (2024), we set λ to 0.48.

SVG, a continuously adjustable reactive power compensation device, is modeled as follows:

$$q^{\text{SVG,min}} \leq q_i^{\text{SVG}} \leq q^{\text{SVG,max}} \tag{6}$$

where $q^{\text{SVG,min}}$ and $q^{\text{SVG,max}}$ are the lower and upper limits of the reactive power output of SVG, respectively.

2.4. Offline-Online Two-Stage Optimization Model

The motivation for introducing the two-stage model is to coordinate the control devices on different timescales. The online regulation of reactive power is dependent on the offline decisions. On the other hand, the offline decision should consider the possible online adjustment of reactive power. In other words, the feasible region of offline variable is interdependent with that of online ones. For example, a transformer tap can mitigate overvoltage during peak PV hours but may lead to under-voltage in other times.

The proposed approach is to improve the overall voltage quality in months. Therefore, the load profiles can be significantly different, and uncertainties may be introduced in the model Zhang et al. (2025). In this work, we focus on the practical feasibility and employ a scenario-based approach to handle the uncertainty. S scenarios are generated with the same probability $1/S$ Zhao and Guan (2013). The offline stage is to decide low-voltage transformer taps and the control policy for low-voltage capacitors. The equipment cannot be controlled remotely in real-time. They are adjusted every few months. The online stage is to adjust the PV and SVG. It is noted that the reactive power of capacitor also changes based on the control policy in the online stage. Let t and s denote time and scenario indices, respectively. Let $\mathcal{L}$ denote the set of regular lines, and $\mathcal{T}$ denote the set of equivalent line for transformers. We formulate the scenario-based optimization model as follows.

$$\text{(P)} \quad \min \quad \frac{1}{S}\sum_{s\in\mathcal{S}}\sum_{t\in T}\left(\mathbf{C}^{V}\cdot\sum_{i\in I}\left(\bar{\zeta}_{its}+\underline{\zeta}_{its}\right)+\mathbf{C}^{\mathrm{curt}}P^{\mathrm{curt}}_{its}\right) \tag{7}$$

$$\text{s.t.} \quad \sum_{k:i\to k} P_{kts} = P_{its} - R_i\frac{P_{its}^2+Q_{its}^2}{V_{its}^2} + p_{its} \tag{8}$$

$$\sum_{k:i\to k} Q_{kts} = Q_{its} - X_i\frac{P_{its}^2+Q_{its}^2}{V_{its}^2} + q_{its} \tag{9}$$

$$V_{kts}^2 = V_{its}^2 - 2(P_{its}R_i + Q_{its}X_i) + (R_i^2+X_i^2)\frac{P_{its}^2+Q_{its}^2}{V_{its}^2}, \quad (i\to k)\in\mathcal{L} \tag{10}$$

$$V_{kts}^2 = \frac{V_{its}^2}{K_i^2} - 2(P_{its}R_i + Q_{its}X_i) + K_i^2(R_i^2+X_i^2)\frac{P_{its}^2+Q_{its}^2}{V_{its}^2}, \quad (i\to k)\in\mathcal{T} \tag{11}$$

$$p_{its} = p_{its}^{\mathrm{pv,actual}} - p_{its}^{\mathrm{curt}} - p_{its}^{\mathrm{load}} \tag{12}$$

$$q_{its} = q_{its}^{\mathrm{pv}} - q_{its}^{\mathrm{load}} + q_{its}^{\mathrm{SVG}} + q_{its}^{\mathrm{CB}} \tag{13}$$

$$q^{\mathrm{SVG,min}} \le q_{its}^{\mathrm{SVG}} \le q^{\mathrm{SVG,max}} \tag{14}$$

$$q_{its}^{\mathrm{CB}} = y_{its}\cdot q_i^{\mathrm{CB,com}} \tag{15}$$

$$0 \le y_{its} \le y^{\max} \tag{16}$$

$$y_{its} \in Z \tag{17}$$

$$V_{its} - \bar{\zeta}_{its} \le V^{\max}, \quad V_{its} + \underline{\zeta}_{its} \ge V^{\min} \tag{18}$$

$$\bar{\zeta}_{its} \geq 0, \quad \underline{\zeta}_{its} \geq 0 \tag{19}$$

$$0 \leq p_{its}^{\text{curt}} \leq p_{its}^{\text{pv,actual}} \tag{20}$$

$$-0.48 p_{its}^{\text{pv,actual}} \leq q_{its}^{\text{pv}} \leq 0.48 p_{its}^{\text{pv,actual}} \tag{21}$$

$$\frac{p_{its}^2}{p_{its}^2 + q_{its}^2} \geq \cos\phi \tag{22}$$

The objective is to minimize PV curtailment and mitigate voltage violations. As a side note, the objective function can easily be modified, such as minimizing power loss or cost. Decision variables include transformer ratio K_i, PV reactive power q_{its}^{pv}, PV real power curtailment p_{its}^{curt}, SVG reactive power q_{its}^{SVG}, and capacitor reactive power q_{its}^{CB} and PV reactive power p_{its}. Constraints (8-11) denote the power flow model. The real and reactive power injections are represented by equations (12) and (13), respectively. The SVG reactive power is limited by (14). The capacitor banks are modeled in (15) and (16). The voltage limit is enforced in (18). The reactive power of PV generator is limited by (21). Equation (22) denotes the power factor constraint.

Equation (22) can be recast as

$$p_{its} \geq \frac{\cos\phi}{\sqrt{1 - \cos\phi^2}} \cdot q_{its} \tag{23}$$

Next, a data-driven approach will be introduced to solve the mixed integer nonlinear programming problem (P).

3. Data-Driven Models

3.1. Linear Models for Power Flow and Transformers

In problem (P), constraints (8)(9)(10)(11) are nonlinear. In particular, integer variables are included in the constraint (11). After comprehensive experiments, we find that the multiple linear regression model can accurately approximate the nonlinear power flow. The linear DistFlow model Baran and Wu (1989), which relies on the zero-loss assumption, can be considered an early form of a data-driven approach, as this simplification is justified by typical voltage data. This work further extends it with better accuracy and includes integer variables for transformer tap. A multiple linear regression model can be expressed as follows

$$y = b_1x_1 + b_2x_2 + \cdots + b_nx_n + b_{n+1} \tag{24}$$

where $x_1, x_2, \cdots, x_n$ are the independent variables, y is the dependent variable, and $b_1, b_2, \cdots, b_{n+1}$ are the linear regression coefficients. The coefficients of the linear regression can be determined using the least squares method. The regression coefficients $b_1, b_2, \cdots, b_{n+1}$ are obtained by solving the normal equations using the least squares method.

The relationship between node voltage magnitudes and reactive/active power can be described by a multiple linear regression formulation as follows:

$$\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} = \begin{bmatrix} C_1 & A_{11} & \cdots & A_{1n} & B_{11} & B_{12} & \cdots & B_{1n} \\ C_2 & A_{21} & \cdots & A_{2n} & B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ C_n & A_{n1} & \cdots & A_{nn} & B_{n1} & B_{n2} & \cdots & B_{nn} \end{bmatrix} \begin{bmatrix} 1 \\ p_1 \\ \vdots \\ p_n \\ q_1 \\ \vdots \\ q_n \end{bmatrix} \tag{25}$$

where $B_{i,j}$ is the regression coefficient quantifying the sensitivity of V_i to q_j, and A_{ij} is the regression coefficient quantifying the sensitivity of V_i to p_j, and c_i is the intercept term that captures the nominal voltage component independent of q_j and p_j. These coefficients are identified using historical operating data through least squares regression analysis.

Let $\mathcal{I}_x$ denote the set of all nodes on the low voltage side of the transformer x. The voltage for nodes in $\mathcal{I}_x$ can be expressed by a multiple linear regression formulation

$$V_i = \sum_{k=1}^{K} C_{ik}\beta_{xk} + \sum_{j=1}^{|\mathcal{I}_x|} A_{ij}p_j + \sum_{j=1}^{|\mathcal{I}_x|} B_{ij}q_j, \quad i \in \mathcal{I}_x \tag{26}$$

where β_{xk} is the binary indicator for transformer tap k. K presents the number of tap positions.

The problem (P) can be recast as

$$\begin{aligned}
\text{(Lin-P)} \quad \min \quad & \frac{1}{S}\sum_{s\in\mathcal{S}}\sum_{t\in T}\left(\mathbf{C}^{V}\cdot\sum_{i\in I}\left(\overline{\zeta}_{its}+\underline{\zeta}_{its}\right)+\mathbf{C}^{\text{curt}}P^{\text{curt}}_{its}\right) & \\
\text{s.t.} \quad & \sum_{k:i\rightarrow k} P_{kts} = P_{its} + p_{its}, & (27)\\
& \sum_{k:i\rightarrow k} Q_{kts} = Q_{its} + q_{its} & (28)\\
& (25)(26)(12) - (21)(23) &
\end{aligned}$$

The coefficient matrices can be found according to Figure 4. The flow can be divided into three main phases: dataset preparation, model training, and model testing.

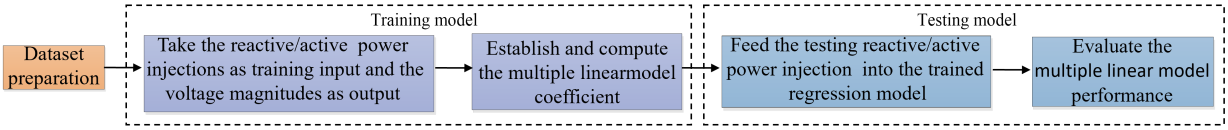

Dataset preparation
Training model
Take the reactive/active power injections as training input and the voltage magnitudes as output
Establish and compute the multiple linearmodel coefficient
Testing model
Feed the testing reactive/active power injection into the trained regression model
Evaluate the multiple linear model performance

Figure 4: Flow chart of the learning-based power flow linearization.

3.2. *Learning Control Policy of Capacitor Banks*

The optimal control y^*_{its} at node i and time t in scenario s can be generated by solving the problem (Lin-P). However, many low-voltage capacitor banks do not have communication functions and cannot be dispatched in real time. In other words, y^*_{its} is meaningless for these capacitor banks. We propose training a machine learning model and implementing it offline. In the online stage, the capacitor is triggered by local information. We employ the Random Forest to predict y_{its}. The flow chart to predict capacitor steps are outlined in Figure 5.

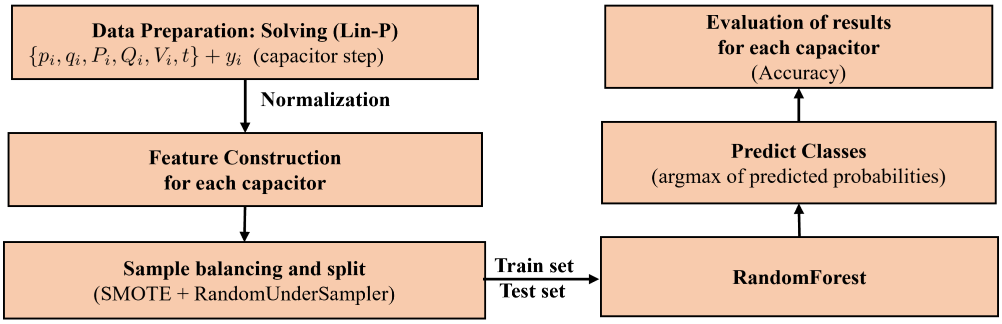
- The process begins with data preparation, where real power injection p_i, reactive power injection q_i, real power flow P_i, reactive power flow Q_i, voltage magnitude V_i, and the time index (t, representing the hour of the day from 0 to 23) are collected together with the capacitor step y_i labels. These data are generated by solving problem (Lin-P).

- After normalization, features are constructed separately for each capacitor.

- Most of the time, there is no need to dispatch additional reactive power. A challenge is hence the imbalance of the data set. To mitigate class imbalance, a hybrid resampling strategy combining SMOTE and RandomUnderSampler is applied to the data set after the train-test split Yang et al. (2024); Chawla et al. (2002).

- A Random Forest model is then trained, and class labels are assigned directly by selecting the class with the highest predicted probability.

- Finally, the prediction performance for each capacitor is evaluated using accuracy and related metrics.



Data Preparation: Solving (Lin-P)
$\{p_i, q_i, P_i, Q_i, V_i, t\} + y_i$ (capacitor step)
Normalization
Feature Construction for each capacitor
Sample balancing and split (SMOTE + RandomUnderSampler)
Train set
Test set
RandomForest
Predict Classes (argmax of predicted probabilities)
Evaluation of results for each capacitor (Accuracy)

Figure 5: Flow chart of learning control policy for capacitor banks.

4. Case Study

The simulation is performed with a real-world distribution network in Suqian, China. The network comprises 520 nodes, including 122 medium-voltage nodes and 398 low-voltage nodes. The base voltages for the medium- and low-voltages are 10 kV and 0.38 kV, respectively. For clarity, Figure 6 illustrates only the 122 medium-voltage nodes, and the low-voltage nodes are situated within the red-highlighted substation areas. In total, 26 PV units are considered in this study. As shown in Figure 6, three PV units are connected to medium-voltage nodes, with a total installed capacity of 7,640 kW. The remaining 23 PV units are distributed across four low-voltage distribution areas with a total capacity of 693.58 kW. The PV deployment is summarized in Table 1.

Table 1: Number of PV in the distribution network

Distribution	Node number	PV number	PV installed capacity/kW
Medium voltage feeders	122	3	7640.00
LV distribution area 1	62	6	193.47
LV distribution area 2	149	8	188.50
LV distribution area 3	76	4	150.15
LV distribution area 4	111	5	161.46

The data are collected from 00:00 on June 1, 2024, to 23:00 on June 27, 2024, with a 15-minute sampling interval. 27 representative scenarios for summer are selected from historical annual data. Each scenario consists of 96 periods (96=24*4). The case study is implemented in MATLAB R2024a and solved by Gurobi solver. All simulations are executed on an Intel Core i5-3210 at 3.5 GHz.

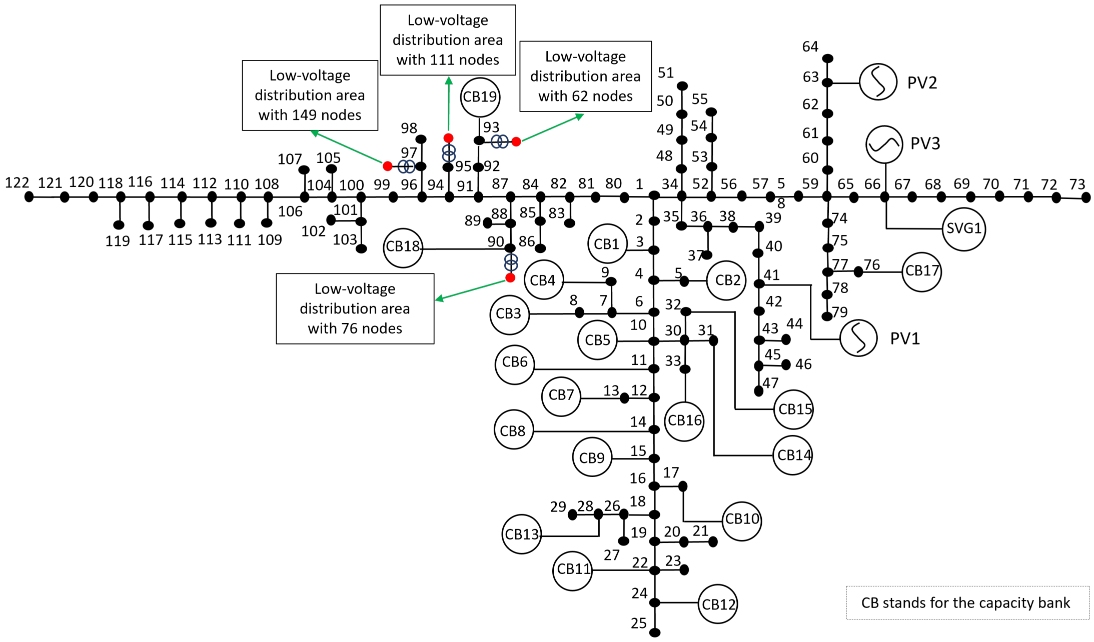

Low-voltage distribution area with 111 nodes
Low-voltage distribution area with 149 nodes
Low-voltage distribution area with 62 nodes
Low-voltage distribution area with 76 nodes
CB19
CB18
CB1
CB2
CB3
CB4
CB5
CB6
CB7
CB8
CB9
CB10
CB11
CB12
CB13
CB14
CB15
CB16
CB17
PV1
PV2
PV3
SVG1
CB stands for the capacity bank

Figure 6: Schematic diagram of the real-world distribution system

4.1. Performance of Voltage Violation Mitigation

To verify the effectiveness of the proposed method and analyze voltage violations, four methods are used in the simulations.

- Method 1: The current heuristic method used in practice.
- Method 2: Offline optimization, optimizing transformer tap.
- Method 3: Online optimization, optimizing PV inverters and SVG reactive power.
- Method 4: Offline and online optimization, optimizing transformer tap and control policy offline, and optimizing PV inverters and SVG reactive power.

In this case study, the simulation covers 648 hours with 27 scenarios, yielding a total of 336,960 periods. The total available PV generation in these periods amounts to 20.815 MWh and the Photovoltaic installed capacity is 693.58 kW. The key performance metrics are voltage violation rate and curtailment rate. They are used to assess the effectiveness of the optimization results. The voltage violation rate is defined as the ratio of the number of voltage magnitude data exceeding the limit to the total number of observation points. The curtailment rate is defined as the ratio of curtailed PV generation to the total available PV energy.

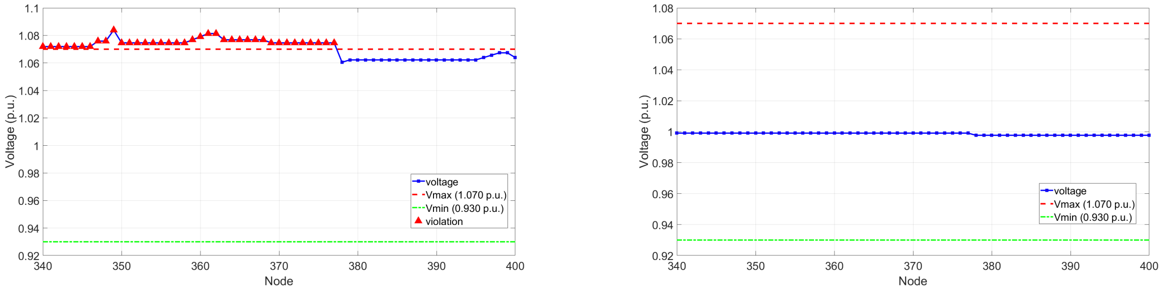
4.1.1. Results with base PV capacity

Table 2 summarizes the results for all four methods. Column “Method 1" is the data collected from the real system. There is 3.672% voltage violation. All optimization-based methods (Methods 2–4) successfully eliminate voltage violations.

Table 2: Violation comparison of optimization methods

Method	Method 1	Method 2	Method 3	Method 4
Available PV Generation/MWh	20.815	20.815	20.815	20.815
Violation Number	12374	0	0	0
Violation Rate	3.672%	0.000%	0.000%	0.000%
Curtailment/MWh	0	0	0	0
Curtailment Rate	0.000%	0.000%	0.000%	0.000%
Transformer Tap	1,2,2,2	1,1,1,1	1,2,2,2	1,1,1,1

The medium-voltage magnitude attained from Method 1 and Method 4 are shown in Figure 7(a) and (b). The x-axis represents the node, while the y-axis denotes the voltage magnitude. As shown in Figure 7, the proposed offline-online optimization method effectively prevents voltage violations. The violation numbers of Method 1 is 12374, while there is no violations in results attained by Method 4.



(a) Violation of Method 1.

(b) Violation of Method 4.

Figure 7: Comparison of optimization methods for violation.

4.1.2. *Results with Varying PV Capacity*

Figure 8 illustrates the PV curtailment results under varying installed PV capacities, with voltage limit enforced. The red, blue, green, and black solid lines correspond to the outcomes of Method 1, Method 2, Method 3, and Method 4, respectively. The x-axis represents the installed PV capacity, ranging from 694 kW to 17,340 kW, while the y-axis indicates the total PV curtailment for each method. As observed in Figure 8, Method 4 achieves the lowest PV curtailment across all capacity levels, without any voltage violations. Furthermore, the performance advantage of Method 4 becomes increasingly pronounced as the installed PV capacity rises. For instance, at the capacity of 10,404 kVA, the PV curtailment from Method 1 is approximately 73 MWh, whereas Method 4 limits curtailment to only 0.9 MWh. These results demonstrate that the proposed offline-online voltage regulation strategy significantly mitigates PV curtailment respecting voltage constraints, particularly in scenarios with high PV penetration.

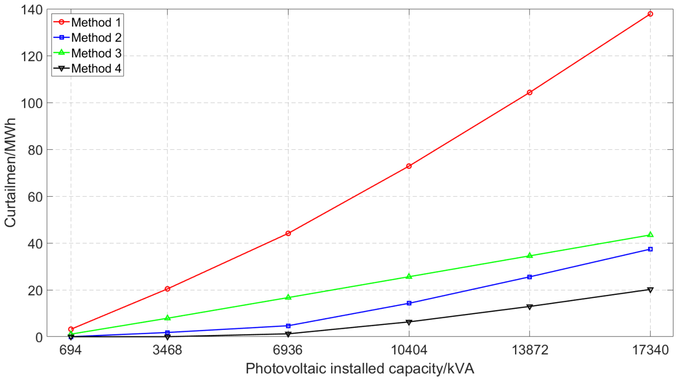


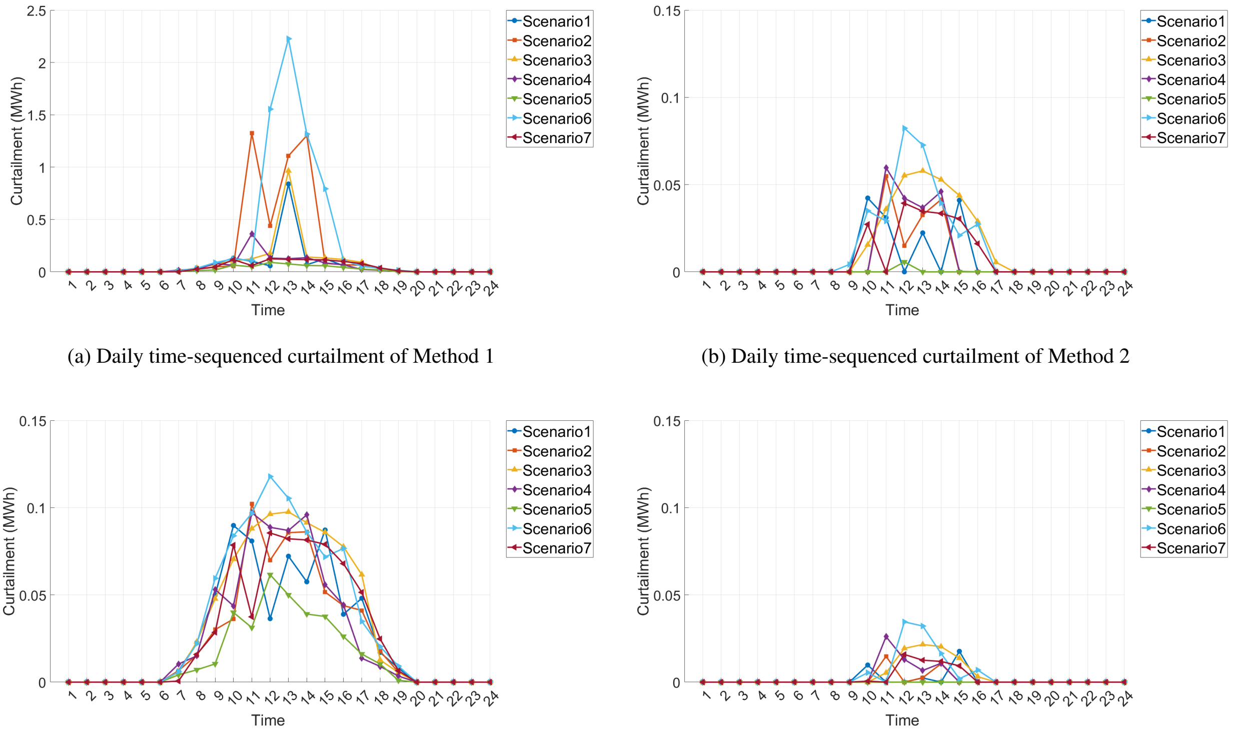
Figure 8: Optimization methods comparison of curtailment under different PV capacity

Table 3 presents a comparison of four optimization methods after scaling the installed PV capacity by a factor of 10. The energy generated by distributed PV is 208.153 MWh. The voltage limit is enforced for all methods by allowing PV curtailment. The transformer tap positions are (1,1,1,1) after optimization. Method 1 results in a PV curtailment of 44.121 MWh, corresponding to a curtailment rate of 21.196%. In contrast, Method 4 achieves a significantly lower curtailment of only 1.232 MWh, or 0.592%. This demonstrates that the proposed offline-online coordinated optimization framework (Method 4) yields the most favorable outcome, reducing the PV curtailment rate by up to 20.604 percentage points compared to the baseline (Method 1).

Table 3: Comparison of optimization methods

Method	Method 1	Method 2	Method 3	Method 4
Available PV Generation/MWh	208.153	208.153	208.153	208.153
Violation Number	0	0	0	0
Violation Rate	0.000%	0.000%	0.000%	0.000%
Curtailment/MWh	44.121	4.683	16.708	1.232
Curtailment Rate	21.196%	2.250%	8.027%	0.592%
Transformer Tap	2,2,2,2	1,1,1,1	2,2,2,2	1,1,1,1

Figure 9 presents the PV curtailment results obtained by four methods in seven representative scenarios. The daily time-sequenced curtailment results of Methods 1 and 2 are shown in Figure 9(a) and 9 (b). The daily time-sequenced curtailment results of Methods 3 and 4 are shown in Figure 9(c) and 9 (d). It is observed that method 4 achieves the lowest curtailment in all seven scenarios. The largest PV curtailment by Method 1 observed at Hour 13 in Scenario 6, amounting to 2.2 MWh. This scenario characterized by sunny weather, leading to a high available PV generation of 5.54 MWh between 13:00 and 14:00. Additionally, being a Saturday, the load demand is relatively low. Therefore, the highest PV curtailment occurs in Scenario 6. In contrast, the largest curtailments from Methods 2 and 3 are 0.082 MWh and 0.12 MWh, respectively. Method 4 attains the least curtailment of 0.035 MWh.



(a) Daily time-sequenced curtailment of Method 1

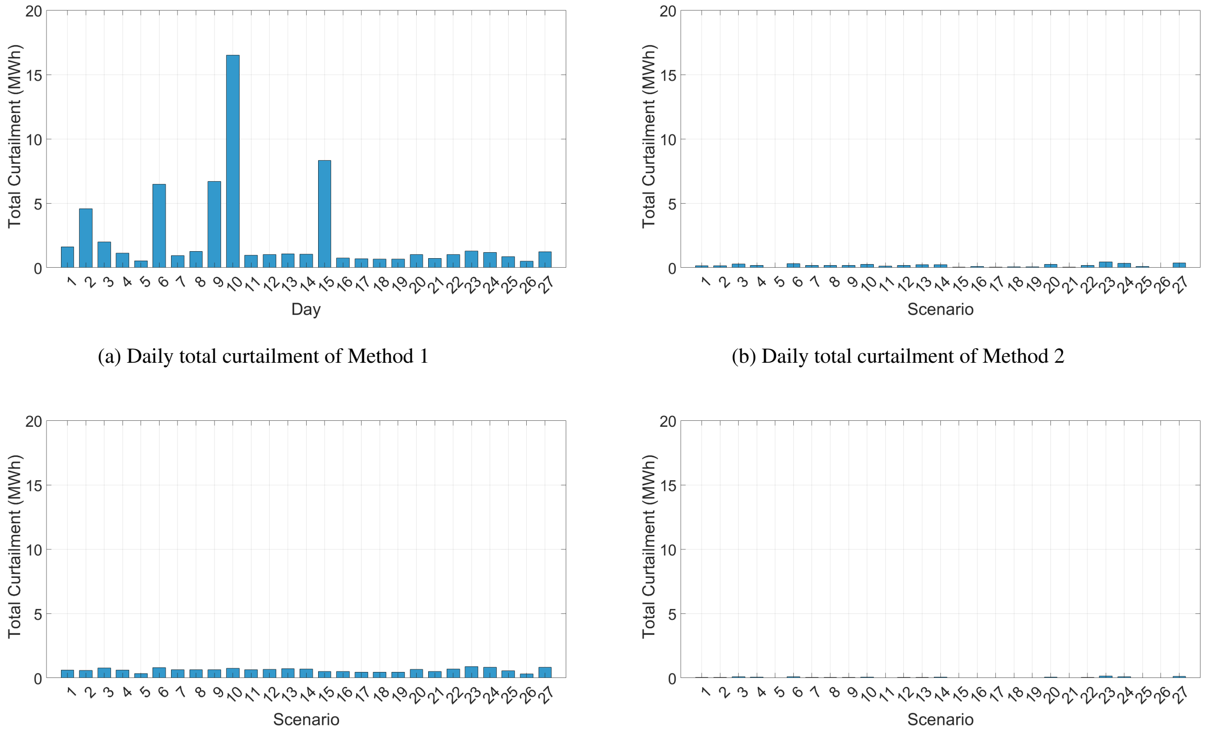
(b) Daily time-sequenced curtailment of Method 2

(c) Daily time-sequenced curtailment of Method 3

(d) Daily time-sequenced curtailment of Method 4

Figure 9: Daily time-sequenced curtailment results comparison of different methods. For readability, seven typical scenarios are selected.

Figure 10 presents the daily total PV curtailment results attained by four methods. Subfigures (a) and (b) correspond to Methods 1 and 2, while subfigures (c) and (d) display the results for Methods 3 and 4, respectively. As illustrated in Figure 10 and summarized in Table 3, both the absolute curtailment and curtailment rates differ significantly among the methods. Specifically, Method 1 yields a total curtailment of 44.121 MWh, corresponding to a curtailment rate of 21.196%. Method 2 reduces the curtailment to 4.683 MWh (2.250%), while Method 3 results in 16.708 MWh (8.027%). The proposed offline-online coordinated optimization (Method 4) achieves the lowest curtailment of 1.232 MWh, with a curtailment rate of only 0.592%. These results demonstrate the superior performance of the proposed approach in minimizing PV curtailment with voltage constraints.



(a) Daily total curtailment of Method 1

(b) Daily total curtailment of Method 2

(c) Daily total curtailment of Method 3

(d) Daily total curtailment of Method 4

Figure 10: Daily total curtailment results comparison of different methods.

4.2. Data-Driven Linear Approximation

The computational performance of the proposed method is compared with the second order cone programming (SOCP) based approach, which is widely used in the literature. The solution time, model size, and memory usage are presented in Table 4. It can be observed that SOCP-based approach finds the solution for the 24 hour cases with more than six minutes. In contrast, the data-driven method obtains the solution in 1.05 seconds. When the time horizon extends to 48 hours, the SOCP-based method requires over 40 minutes (2506.18 seconds) to get the solution, while the data-driven method only takes 2.59 seconds. The results demonstrate that the proposed data-driven method significantly reduces computational time.

Table 4: Multi-period comparison between the SOCP and data-driven method

Method	Single Period		24 Periods		48 Periods	
	SOCP	Data-driven	SOCP	Data-driven	SOCP	Data-driven
Solution Time (seconds)	83.83	0.15	377.47	1.05	2506.18	2.59
# of Linear Constraints	3193	3645	76632	87388	153264	174772
# of Nonlinear Constraints	519	0	12456	0	24912	0
# of Variables	6835	4773	164040	114092	328080	228164
Violation Rate (%)	0.00	0.00	0.00	0.00	0.00	0.00
Memory Usage (MB)	4354.38	3923.51	4987.34	3983.12	5214.88	4104.13

4.2.1. Training Performance with Various Dataset

In this section, we evaluate the accuracy of the multiple linear regression-based power flow model. All voltage magnitudes, active power, and reactive power values are converted to per-unit (p.u.) values during preprocessing. The model utilizes the active and reactive power injections at all nodes as input features, with the corresponding nodal voltage magnitudes serving as output targets. Input and output datasets are generated from power flow calculations, and the data are subsequently partitioned into training and testing sets, with the testing set comprising 25% of the total samples.

Table 5: Voltage accuracy of low-voltage nodes along with increasing size of data set

# of Training Samples	**# of Test Samples**	**Training RMSE (p.u.)**	**Test RMSE (p.u.)**
200	50	1.46×10^{-4}	2.29×10^{-4}
300	75	1.71×10^{-4}	2.16×10^{-4}
400	100	1.75×10^{-4}	2.22×10^{-4}
500	125	1.73×10^{-4}	2.09×10^{-4}
600	150	1.67×10^{-4}	2.09×10^{-4}
700	175	1.71×10^{-4}	2.14×10^{-4}
800	200	1.80×10^{-4}	2.10×10^{-4}

Tables 5 and 6 summarize the root mean squared error (RMSE) of voltage predictions for low-voltage and medium-voltage nodes, respectively, as a function of training set size. The column "# of Training Samples" indicates the number of samples used for model training, while "RMSE" reports the corresponding prediction error.

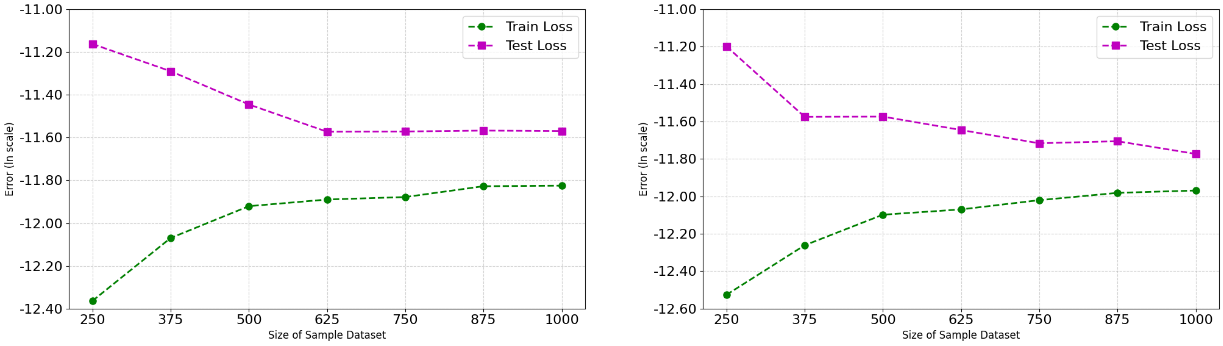
For low-voltage nodes, the overall RMSE remains on the order of 10^{-4} p.u. For instance, with a training set size of 400, the training RMSE is 1.75×10^{-4} p.u. (Table 5). For medium-voltage nodes, the RMSE is even lower, typically in the range of 10^{-5} to 10^{-6} p.u.; for example, with 600 training samples, the test RMSE is 9.43×10^{-6} p.u. (Table 6).

The results further indicate that model accuracy improves as the training dataset size increases, reaching optimal performance at 500 training samples, where the lowest RMSEs are 2.09×10^{-4} p.u. for low-voltage nodes and 9.42×10^{-6} p.u. for medium-voltage nodes. These findings demonstrate that the proposed data-driven linearization method enables highly accurate voltage estimation for both low- and medium-voltage nodes in practical distribution networks.

Table 6: Voltage accuracy of middle-voltage nodes along with increasing size of data set

# of Training Samples	**# of Test Samples**	**Training RMSE (p.u.)**	**Test RMSE (p.u.)**
200	50	4.27×10^{-6}	1.42×10^{-5}
300	75	5.73×10^{-6}	1.25×10^{-5}
400	100	6.65×10^{-6}	1.07×10^{-5}
500	125	6.86×10^{-6}	9.42×10^{-6}
600	150	6.94×10^{-6}	9.43×10^{-6}
700	175	7.30×10^{-6}	9.47×10^{-6}
800	200	7.32×10^{-6}	9.45×10^{-6}

Next, we show the impact of the sample range on the data-driven method. The voltage sampling ranges are [0.93, 1.07] and [0.96, 10.7] for Figure 11(a) and (b), respectively. In Figure 11, green dashed circles indicate the logarithmic RMSE of the training samples, while purple dashed squares indicate the logarithmic RMSE of the testing samples. The results in Figure 11 (a) demonstrate that as the sample size increases, the training RMSE gradually increases and stabilizes, while the testing RMSE rapidly decreases and stabilizes. Meanwhile, the small discrepancy between the training error and the testing error across different sample sizes indicates that there is no overfitting. The voltage sampling range for the above results is [0.93, 1.07]. After narrowing this sampling range to [0.96, 1.07], the training and testing procedures are repeated as described above. Figure 11 (b) displays the training and testing RMSE results after narrowing the voltage sampling interval. As shown in Figure 11 (b), reducing the sampling interval yields higher accuracy.



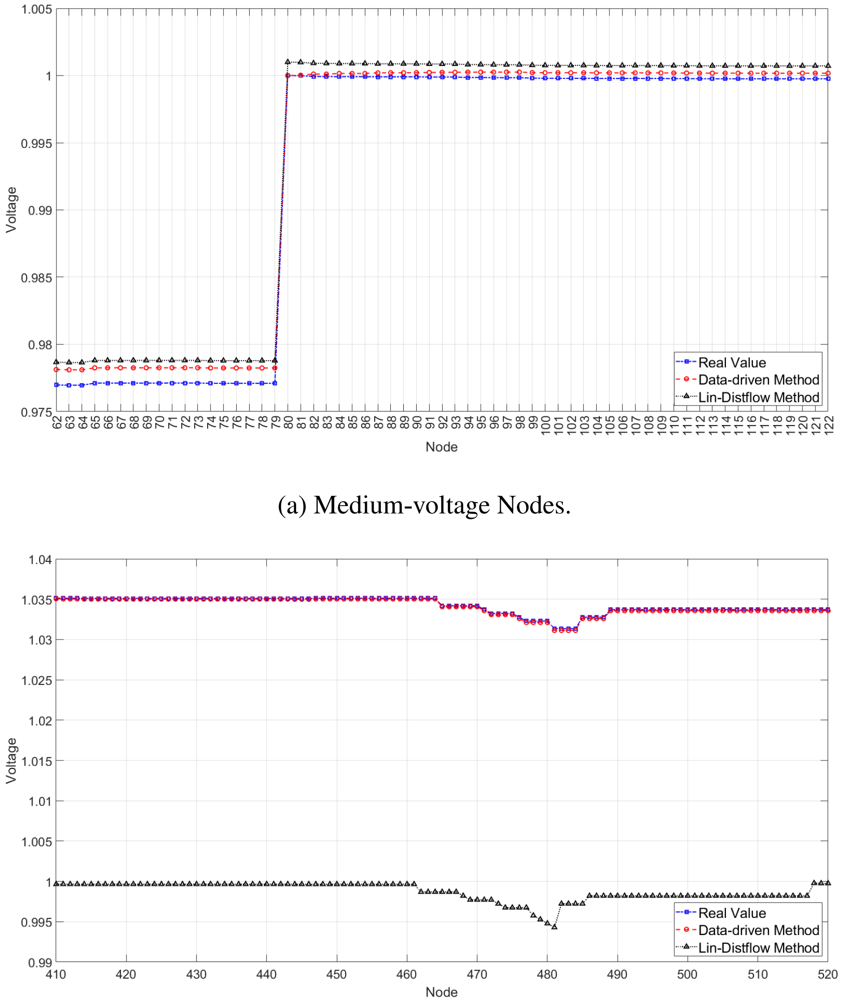
(a) Normal Voltage Range [0.93, 1.07].

(b) Narrowing Voltage Range [0.96, 1.07].

Figure 11: Sample range's impact on accuracy.

4.2.2. *Data-driven Method vs. Lin-DistFlow*

The accuracy of the proposed data-driven method is compared with Lin-DistFlow. The medium voltage RMSE of the proposed method is 9.10×10^{-4}, while that of the Lin-DistFlow is 1.50×10^{-3}. The substation voltage RMSE of the proposed method is 1.28×10^{-5}, while that of the Lin-DistFlow is 3.55×10^{-2}. Figure 12 provides a detailed comparison of voltage estimation accuracy between the proposed data-driven approach and the conventional Lin-DistFlow method. Subfigure (a) displays the results for medium-voltage nodes, while subfigure (b) focuses on low-voltage nodes within a representative substation. In both subfigures, the red circles correspond to the data-driven method, the black triangles denote the Lin-DistFlow results, and the blue squares indicate the actual voltage values obtained from power flow calculations. The data-driven approach consistently provides voltage estimates that closely align with actual values across all nodes, surpassing the accuracy of the Lin-DistFlow method. In particular, for medium-voltage nodes (80–100) and low-voltage nodes (410–520), Lin-DistFlow shows notable deviations from true voltages, while the data-driven method maintains high fidelity. These results confirm the superior accuracy and robustness of the proposed data-driven linearization for modeling voltage profiles in practical distribution networks.



(a) Medium-voltage Nodes.

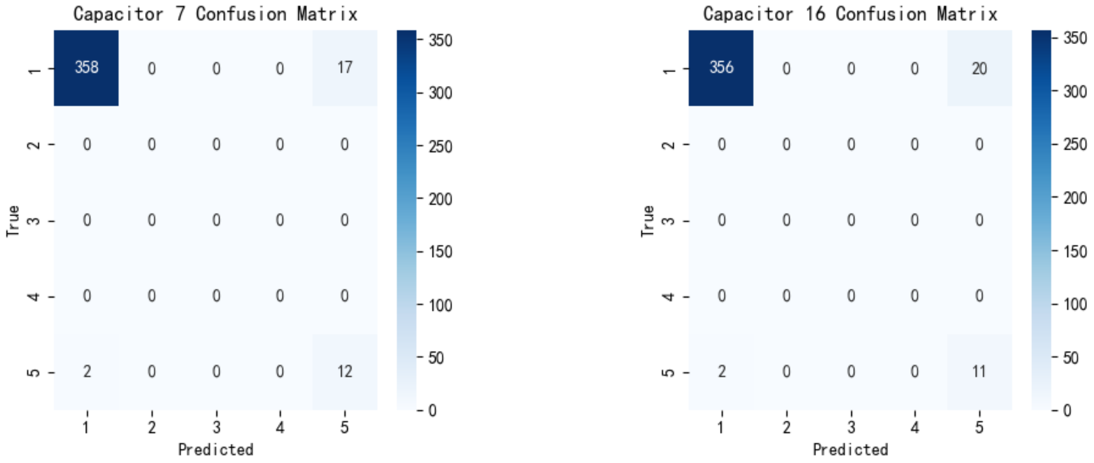
(b) Low-voltage Nodes. Only nodes within a single substation are shown due to space limit.

Figure 12: Accuracy comparison of data-driven and Lin-DistFlow models.

4.2.3. *Learning Output for Capacitor Banks*

At last, we show the effectiveness of control policy learning. The confusion matrices of the learned control policy for capacitor banks are shown in Figure 13. The confusion matrix is widely used for classification models, and it is particularly useful for binary and multi-class classification problems. As shown in Figure 13 (a), the recall rate for Step 5 is approximately 85.71%(85.7%

= 2/12), and the recall rate for Step 1 is approximately 95.47%. Similar trends are observed for capacitor 16 in Figure 13 (b). For all 19 capacitors, the recall rates are 92% and 87% for Step 1 and 5, respectively. As a side note, the prediction accuracy is good enough for voltage violation and PV curtailment. The impact of prediction error on voltage and PV curtailment is ignorable.



	1	2	3	4	5
1	358	0	0	0	17
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	2	0	0	0	12

(a) Confusion matrix for Capacitor 7.

	1	2	3	4	5
1	356	0	0	0	20
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	2	0	0	0	11

(b) Confusion matrix for Capacitor 16.

Figure 13: Predication Results of Learned Control Policy for Capacitor Banks.

Figure 14 illustrates the impact of the learned control policy on PV curtailment across varying installed PV capacities. The red and blue bars represent the results obtained using the learned control policy and the optimal control policy, respectively. The figure shows that the PV curtailment achieved by the learned control policy closely matches that of the optimal control policy for all tested capacity levels. This demonstrates that the data-driven approach for capacitor control policy yields near-optimal performance.

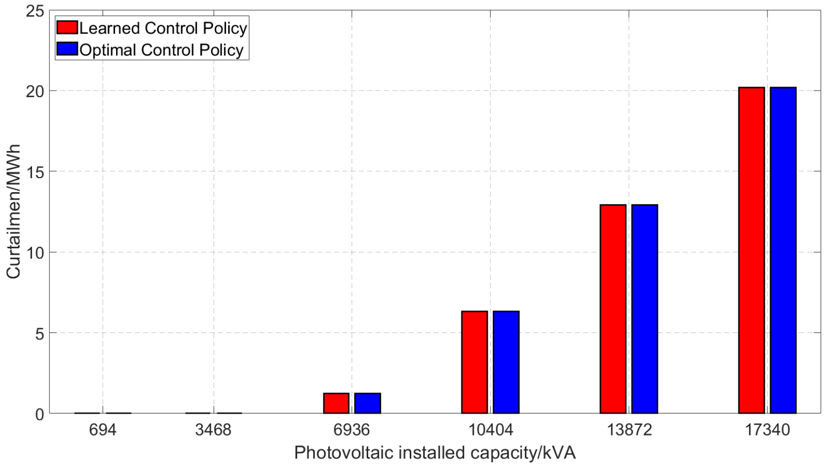


Figure 14: Impact on the PV curtailment with learned control policy.

5. Conclusion

This paper proposes an offline-online voltage optimization framework. By leveraging the various control timescales, this work attains optimal decisions for monthly-changing transformer taps and near-optimal control policies for capacitor banks. The proposed data-driven approach shows good accuracy and computational performance. Random Forest is employed to find the near-optimal control policy for the capacitor banks. It reduces over-voltage rate by 3.672% in a real-world system. It can also decrease PV curtailment from 21.20% to 0.592%. The data-driven linearized power flow model has voltage RMSE of 10^{-6} level. The learned control policy's recall rate can be up to 95%. In future work, we will consider network loss and the energy storage.

6. Acknowledgment

This work is supported by State Grid Jiangsu Science and Technology Projects under project number J2024067.

References

Administration, N.E., 2025. Operation status of renewable energy grid integration in 2024. URL: https://www.nea.gov.cn/20250221/e10f363cabe3458aaf78ba4558970054/c.html. accessed: 2026-01-24.

Agency, I.R.E., 2023. Tripling renewable power and doubling energy efficiency by 2030: Crucial steps towards 1.5°c. URL: https://www.irena.org/Digital-Report/Tripling-renewable-power-and-doubling-energy-efficiency-by-2030. accessed: 2026-01-28.

the April 28 Electricity Crisis, C.A., 2025. The analysis of the electricity crisis of april 28th, 2025. URL: https://www.miteco.gob.es/es/prensa/ultimas-noticias/2025/junio/se-presenta-el-informe-del-comite-de-analisis-de-la-crisis-elect.html.

Baran, M., Wu, F., 1989. Optimal sizing of capacitors placed on a radial distribution system. IEEE Transactions on Power Delivery 4, 735–743. doi:10.1109/61.19266.

Byeon, G., Ryu, M., Kim, K., 2024. Linearized optimal power flow for multiphase radial networks with delta connections. Electric Power Systems Research 235, 110689. URL: https://www.sciencedirect.com/science/article/pii/S0378779624005753, doi:https://doi.org/10.1016/j.epsr.2024.110689.

Carvalho, W.C.D., Ratnam, E.L., Blackhall, L., et al., 2022. Optimization-based operation of distribution grids with residential battery storage: Assessing utility and customer benefits. IEEE Transactions on Power Systems 38, 218–228.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357. doi:10.1613/jair.953.

Chowdhury, M.M.U.T., Murari, K., Hasan, M.S., Kamalasadan, S., 2025. Optimal power flow (opf) analysis for ac–dc active distribution networks utilizing second-order cone programming (socp) approach. IEEE Transactions on Industrial Informatics 21, 3555–3564. doi:10.1109/TII.2025.3528533.

Correa, H.P., Vieira, F.H.T., 2024. Hybrid cooperative markov-based method for decentralized voltage control and loss reduction via photovoltaic reactive power support. Electric Power Systems Research 232, 110398.

Duan, J., Shi, D., Diao, R., Li, H., Wang, Z., Zhang, B., Bian, D., Yi, Z., 2020. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. IEEE Transactions on Power Systems 35, 814–817. doi:10.1109/TPWRS.2019.2941134.

Hong, T., Zhang, Y., 2022. Data-driven optimization framework for voltage regulation in distribution systems. IEEE Transactions on Power Delivery 37, 1344–1347. doi:10.1109/TPWRD.2021.3136773.

Hu, D.E., Peng, Y.G., Wei, W., et al., 2022. Multi-timescale deep reinforcement learning for reactive power optimization of distribution network. Proceedings of the CSEE 42, 5034–5045.

Huang, W.J., Zheng, W.Y., Hill, D.J., 2021. Distribution network reconfiguration for short-term voltage stability enhancement: An efficient deep learning approach. IEEE Transactions on Smart Grid 12, 5385–5395.

IRENA, 2025. Renewable capacity statistics 2025. URL: https://www.irena.org/Publications/2025/Mar/Renewable-Capacity-Statistics-2025.

Jabr, R., 2006. Radial distribution load flow using conic programming. IEEE Transactions on Power Systems 21, 1458–1459. doi:10.1109/TPWRS.2006.879234.

Li, Y., Cao, J., Xu, Y., et al., 2024. Deep learning based on transformer architecture for power system short-term voltage stability assessment with class imbalance. Renewable and Sustainable Energy Reviews 189, 113913.

Meng, L., Yang, X., Zhu, J., et al., 2024. Network partition and distributed voltage coordination control strategy of active distribution network system considering photovoltaic uncertainty. Applied Energy 362, 122846.

Nazih, A., El-Ela, A.A.A., Ali, E.S., 2025. Maximizing hosting capacity of renewable energy sources in unbalanced distribution networks using multi-objective optimization approach. Electric Power Systems Research 242, 111458.

Panel, I.I.E., 2025. Factual Report: Grid Incident in Spain and Portugal on 28 April 2025. Technical Report.

Priyadarshi, R., Kishor, N., Negi, R., Lazzari, R., 2026. Framework for generation scheduling and equivalent dynamic modeling in generation-mix scenarios. Renewable Energy 256, 124333. doi:https://doi.org/10.1016/j.renene.2025.124333.

Standardization Administration of China, 2024. Technical Specifications for Grid Connection of Photovoltaic Power Generation Systems to Distribution Networks. Technical Report GB/T 29319–2024. Standardization Administration of China. Beijing, China. National Standard of the People's Republic of China.

Wang, P., Zhang, Z.Y., Huang, Q., et al., 2021. Robustness-improved method for measurement-based equivalent modeling of active distribution network. IEEE Transactions on Industry Applications 57, 2146–2155.

Wang, S., Zhu, X., Yu, P., et al., 2025. Research on a double-layer voltage and reactive power control strategy for regional distribution networks based on edge computing. Electric Power Systems Research 249, 112052.

Xu, Y., Dong, Z.Y., Zhang, R., Hill, D.J., 2017. Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems. IEEE Transactions on Power Systems 32, 4398–4408. doi:10.1109/TPWRS.2017.2669343.

Yang, C., Fridgeirsson, E.A., Kors, J.A., Reps, J.M., Rijnbeek, P.R., 2024. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. Journal of Big Data 11, 7.

Yeh, H.G., Gayme, D.F., Low, S.H., 2012. Adaptive VAR Control for Distribution Circuits With Photovoltaic Generators. IEEE Transactions on Power Systems 27, 1656–1663. URL: http://ieeexplore.ieee.org/document/6152194/, doi:10.1109/TPWRS.2012.2183151.

Zhang, J., Cui, M., He, Y., 2024. Dual timescales voltages regulation in distribution systems using data-driven and physics-based optimization. IEEE Transactions on Industrial Informatics 20, 1259–1271. doi:10.1109/TII.2023.3274216.

Zhang, L., Ye, H., Ding, F., Li, Z., Shahidehpour, M., 2023. Increasing pv hosting capacity with an adjustable hybrid power flow model. IEEE Transactions on Sustainable Energy 14, 409–422. doi:10.1109/TSTE.2022.3215287.

Zhang, L., Ye, H., Ge, Y., Li, Z., 2025. Ramping-based variable-timescale co-optimization for distribution planning and operation. IEEE Transactions on Power Systems 40, 2519–2531. doi:10.1109/TPWRS.2024.3467276.

Zhao, C., Guan, Y., 2013. Unified Stochastic and Robust Unit Commitment. IEEE Transactions on Power Systems 28, 3353–3361. doi:10.1109/TPWRS.2013.2251916.