Energy Engineering

Tech Science Press

Doi:10.32604/ee.2025.073825

ARTICLE

A Missing Data Complement Method Based on 3D Convolutional Neural Network and CGAN for a Distribution Network

Kewen Li, Xiaoyong Yu, Shifeng Ou* and Jueming Pan

Electric Power Science Research Institute of Guangxi Power Grid Co., Ltd., Nanning, 530023, China

*Corresponding Author: Shifeng Ou. Email: ou_sf.sy@gx.csg.cn

Received: 26 September 2025; Accepted: 25 November 2025; Published: 06 August 2026

ABSTRACT: The increasing integration of renewable energy sources (e.g., wind and solar power) into distribution grids and the development of new, source–grid–load–storage coordinated power systems have led to a substantial expansion in the volume of situational awareness data in the distribution networks. Moreover, the transmission of low-voltage distribution measurement data via a power line carrier (PLC) is often susceptible to packet loss and, consequently, data gaps. To address these issues, this paper proposes a data completion method using a conditional generative adversarial network (CGAN) integrated with a three-dimensional convolutional neural network (3D-CNN). This approach leverages the ability of CNNs to extract and fuse multidimensional spatiotemporal features and the power of GANs (generative adversarial networks) for data augmentation. Firstly, a 3D-CNN is trained to establish a mapping between the spatiotemporal context of the measured data and the target missing data. Secondly, a CGAN is practicing via adversarial training to establish a data completion model for the distribution networks. Finally, the simulations of the IEEE 14-bus and 33-bus systems demonstrate the proposed approach's performance improvement in the distribution networks compared with that of conventional methods in terms of root mean square error, spatiotemporal correlation, and maximum volatility amplitude, which are the typical measuring metrics.

KEYWORDS: Distribution network; data estimation; 3D convolutional neural network; conditional generative adversarial networks; spatiotemporal correlation

1 Introduction

With the continuous integration of distributed energy resources (DERs), such as wind and solar power, into distribution networks, along with the large-scale deployment of source–grid–load–storage coordination systems, the volume of situational awareness data is growing exponentially. Moreover, when massive datasets are transmitted via traditional communication methods, such as mobile networks or power line carriers (PLCs), network congestion and data loss frequently occur. Critically, such data losses could prevent the master station of the distribution automation system from acquiring complete information—including topological parameters and real-time measurements—which in turn compromises system wide situational awareness, undermines dynamic operational control, and disrupts the system's fault-handling and scheduling capabilities [1–3].

Common communication methods in distribution networks include fiber-optic communication, mobile communication, and PLCs. Among these, fiber-optic communication is characterized by high reliability, large bandwidth, and low latency; however, it entails high construction and maintenance costs. In addition, data loss in fiber-optic networks—typically resulting from physical link interruptions, equipment failures, or configuration errors—is often characterized by its sporadic, prolonged, and widespread nature. Mobile communication (e.g., GPRS/3G/4G), another communication method, is susceptible to environmental influences, such as building obstructions, adverse weather conditions, and electromagnetic interference. Data loss in mobile communications is characterized by its intermittent, sudden, and location-dependent nature and is often accompanied by delays or high error rates. In comparison, the PLC provides advantages such as low cost, elimination of wiring, high channel security, and robustness, facilitating its widespread application in distribution networks. However, the narrow bandwidth and complex channel characteristics of PLCs render their data vulnerable to strong noise interference caused by switching operations, nonlinear loads, arcing, channel attenuation, and impedance mismatch. Data loss in PLCs occurs frequently and is characterized by a short duration, strong correlation with grid operational states, and spatiotemporal correlation between adjacent nodes during peak load periods [4]. Therefore, this paper focuses on data loss in PLC-based transmission by investigating corresponding data completion methods.

Currently, missing data in distribution networks are commonly addressed through statistical and machine learning methods. Statistical approaches, such as mean imputation, interpolation, and regression-based methods [5], are valued for their simplicity and interpretability, making them suitable for random missing data caused by human errors or sensor faults. However, they exhibit significant limitations in handling continuous data losses caused by network failures or traffic congestion. In addition, machine learning methods, such as k-nearest neighbor (KNN) [6], random forest (RF) [7], and support vector machine (SVM) [8], leverage predictive modeling to reconstruct missing data and can capture complex nonlinear relationships. However, these methods are parameter sensitive and have limited effectiveness for continuous missing data.

In recent years, deep learning has advanced time-series imputation with architectures that explicitly encode temporal context and long-range dependencies. Recurrent methods such as BRITS [9] use bidirectional gating and learned decay to model irregular sampling and latent dynamics. Transformer–based models such as SAITS [10] capture long-horizon dependencies and heterogeneous correlations without strict stationarity assumptions. Diffusion-based models such as CSDI [11] provide probabilistic imputations with strong sample quality across diverse missingness patterns. In parallel, Generative adversarial networks (GANs) [12] exhibit strong capabilities in learning high-dimensional data distributions. Through adversarial training between the generator and discriminator, GANs iteratively refine the generated samples to approximate the true data distribution [13–16]. Additionally, convolutional neural networks (CNNs) excel at feature extraction and generalization. Recently, their combination has become a prominent research direction in data augmentation. In research on one-dimensional data completion methods, Ref. [17] employed a GAN integrated with an extreme learning machine (ELM) to address missing transient measurement data, while Ref. [18] combined a GAN with a gated recurrent unit (GRU), effectively restoring the temporal continuity of electric vehicle charging load sequences. In addition, in research on multidimensional data completion methods, Ref. [13] developed enhanced data reconstruction frameworks for electrical measurements based on Wasserstein generative adversarial networks (WGANs), Refs. [19,20] developed CGAN-based Generative Adversarial Imputation Nets (GAIN) frameworks, Ref. [21] developed a GAN-based high-frequency component recovery framework to improve the resolution of electrical parameters (e.g., voltage, current, and power), and Ref. [22] developed a bidirectional deep interpolation framework for multisource heterogeneous data. Furthermore, Refs. [23–25] utilized a CNN to extract load temporal characteristics, improving the learning capability of the GRU, and Ref. [26] proposed a method for generating and analyzing power system section diagrams based on high-dimensional data.

However, despite these advances, most generative models focus primarily on temporal characteristics and lack a priori constraints, which could lead to uncontrolled generation, training instability, and nonlinear error accumulation, particularly under high missing data rates over large time scales, while Ref. [27] demonstrated that incorporating conditional information (e.g., voltage, current and phase angle) and associated node data could improve the capture of multidimensional features and increase imputation accuracy. Therefore, in this paper, we propose a 3D-CNN-integrated CGAN model that leverages spatiotemporal feature representation of multimodal power data for missing data completion in distribution networks. Validation conducted on both the IEEE 14-node and IEEE 33-node test systems demonstrates the effectiveness of the proposed approach, which contributes technical support toward observability, measurability, and controllability in distribution networks.

2 A Missing Data Completion Model Based on 3D-CNN+CGAN

To illustrate the rationality of the method proposed in this paper, in this section, the spatiotemporal correlation of distribution network measurement data is first briefly analyzed, and then the construction method of the missing data imputation model based on 3D-CNN+CGAN is discussed from the principle of CGAN.2.1 construction of a 3D data matrix and fusion of spatiotemporal characteristics.

2.1 Spatiotemporal Correlation of Distribution Network Measurement Data

2.1.1 Time Relevance

Residential, commercial, and industrial loads exhibit distinct daily, weekly, and seasonal periodicity. The output of distributed power sources is influenced by meteorological factors such as wind strength and solar irradiance. These factors cause the measured data at distribution network nodes to exhibit autocorrelation and periodicity in time series. Additionally, events such as grid faults, switch operations, and load switching can cause transient changes and temporal correlations in measurement data.

2.1.2 Spatial Correlation

Kirchhoff's current law and voltage law dictate that electrical quantities at adjacent nodes/branches exhibit strong coupling relationships. Changes in the injected power at a node can influence the voltage/current distribution upstream, downstream, and across the entire feeder. Since distribution grids typically adopt a radial or weak ring network structure, nodes with physically proximate locations often exhibit higher correlations in their electrical states (e.g., power and voltage magnitude). Additionally, the load patterns of the feeder trunk lines and branches also exhibit spatial clustering.

From the above analysis, it can be concluded that the physical structure of distribution networks, operational constraints, and spatiotemporal behavioral patterns of the user side and generation side collectively determine the inherent, model-learnable correlations in measurement data across temporal and spatial dimensions.

2.2 CGAN Principles

The principles of generative adversarial networks (GANs) can be summarized as follows: The GAN generator G generates pseudodata to deceive the discriminator D, while the discriminator D strives to distinguish between real data and generated data. When the two undergo adversarial training to reach a Nash equilibrium, the generator G can generate data that conforms to the distribution of real data. The structure of the CGAN is shown in Fig. 1. It adds conditional information c to the GAN structure to form a semi supervised learning framework, thereby strengthening the network training objective. In this paper, the voltage data of the distribution network nodes are used as the conditional variable c, and the power data of neighboring nodes are used as the noise z. The generator G generates power data that satisfy the voltage conditions of the missing nodes; the discriminator D judges the similarity between the generated power data and the real data, outputs the discrimination value, and feeds it back to itself and the generator G, enabling both to continuously compete and optimize their capabilities until the competition reaches equilibrium.

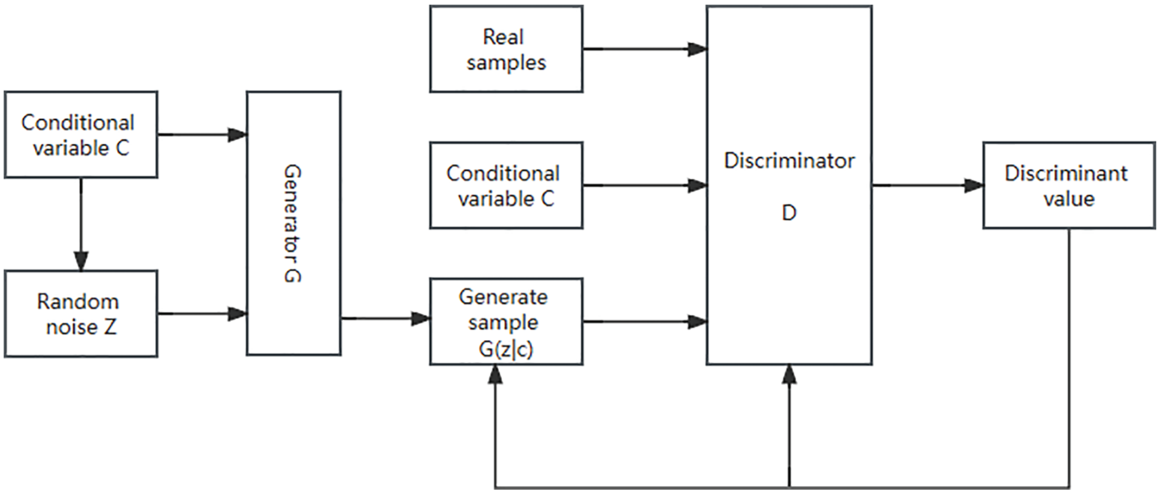

Figure 1: Basic structure of the CGAN

At present, the Conditional Generative Adversarial Network (CGAN) framework has been widely applied in the generation and imputation of various physical time-series data, demonstrating its powerful capability to produce high-quality, conditionally controlled data. Such research validates the effectiveness and generalizability of CGAN models in handling data from complex physical systems. For instance, in the field of energy load forecasting, Sajjad et al. [23] combined CNN with GRU networks and utilized generative models to enhance short-term residential load data, showcasing the ability of generative models to capture the temporal characteristics of user consumption behavior. In environmental and climate science, Zhang et al. [24] developed a model based on generative adversarial networks to simulate and predict building energy consumption sequences under different climatic conditions, successfully learning the complex nonlinear mapping between meteorological parameters and energy usage.

2.3 Spatiotemporal Characteristics of Distribution Network Measurement Data

Assume that a node with missing data has m neighboring nodes. c_i and Z_i represent the voltage data and power data input of the i-th node, respectively, and are used as the conditional input and random noise input for the CGAN. The voltage data c and power data z of the m nodes are fused into m flat matrices. The fused data matrix of the i-th node is shown in Fig. 2.

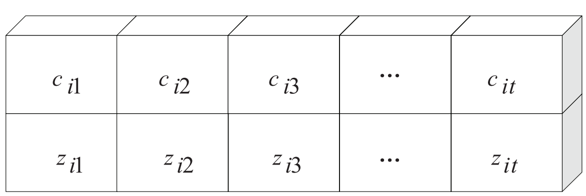

Figure 2: Refactoring the data matrix

Based on Fig. 2, each neighboring node already has a fusion data matrix. At this point, the fusion data matrices of m neighboring nodes are concatenated to form a three-dimensional data matrix, as shown in Fig. 3.

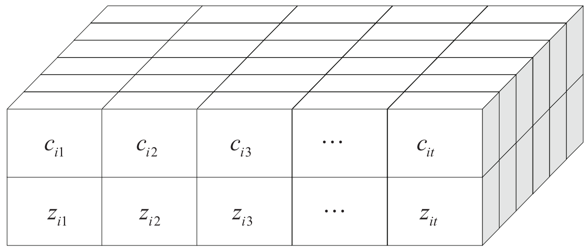

Figure 3: Three-dimensional tensor data structure

Using a 3D convolution kernel (Conv3), the spatiotemporal characteristics of the 3D data matrix in Fig. 3 are extracted, as shown in Fig. 4. The 3D convolution kernel uses a sliding time window on the red-marked plane to extract temporal and spatial features separately. 3D-CNN is a data-driven machine learning method. This method uses local connections, weight sharing, and hierarchical feature extraction characteristics to learn spatiotemporal correlations from three-dimensional measurement data matrices with node-time grid structures to obtain data features. Compared with 2D-CNN, which ignores time parameters, or Recurrent Neural Network (RNN), which has weak spatial relationship modeling capabilities, 3D-CNN achieves higher feature extraction accuracy by using three-dimensional data parameters.

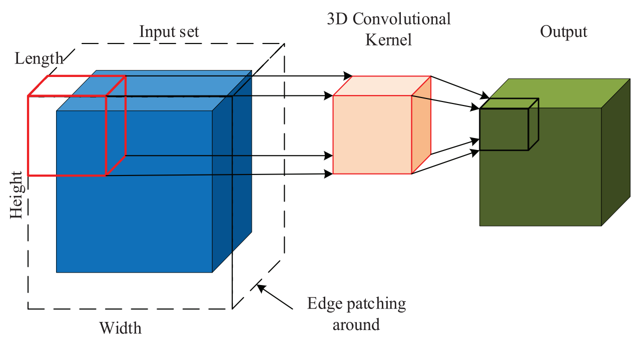

Figure 4: 3D convolution flowchart

2.4 3D-CNN+CGAN Network

Considering the three-dimensional data matrix input in Section 2.3, the 3D-CNN serves as a spatiotemporal feature extractor, mapping spatiotemporal data blocks with missing values into feature vectors rich in spatiotemporal information. By combining a 3D convolutional neural network to enhance the spatiotemporal feature extraction capabilities of the CGAN structure, we obtain the generator network and discriminator network structures of the proposed 3D-CNN+CGAN data completion model, as shown in Figs. 5 and 6.

Considering the possibility of negative values during network training, the generator network consists of 8 layers, and its structure is:

$$INPUT\_layer - [Conv3 - BN - LeakyReLU]_i - OUTPUT\_layer$$

where i represents the i-th intermediate layer, and there are 6 such layers in total. $INPUT\_layer$ represents the input layer; $OUTPUT\_layer$ represents the output layer; $Conv3$ represents 3D convolution; BN represents batch normalization. The structure of the input layer and intermediate layers is $Conv3 - BN - LeakyReLU$, all using the $LeakyReLU$ activation function. The structure of the output layer is $Conv3 - ReLU$, using the ReLU activation function. The 3D convolution kernel size is all $3 \times 3 \times 3$ (with a stride of 1 and padding of 1), and the number of convolution kernels is 64, 128, 256, 512, 256, 128, and 64 in sequence.

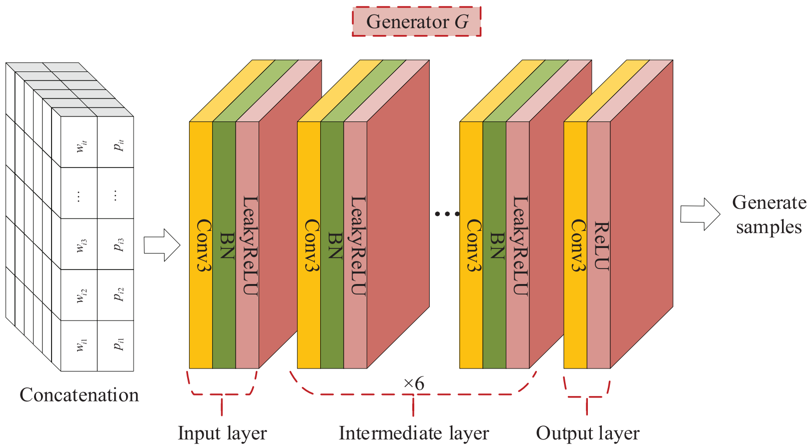

Figure 5: Generator network structure

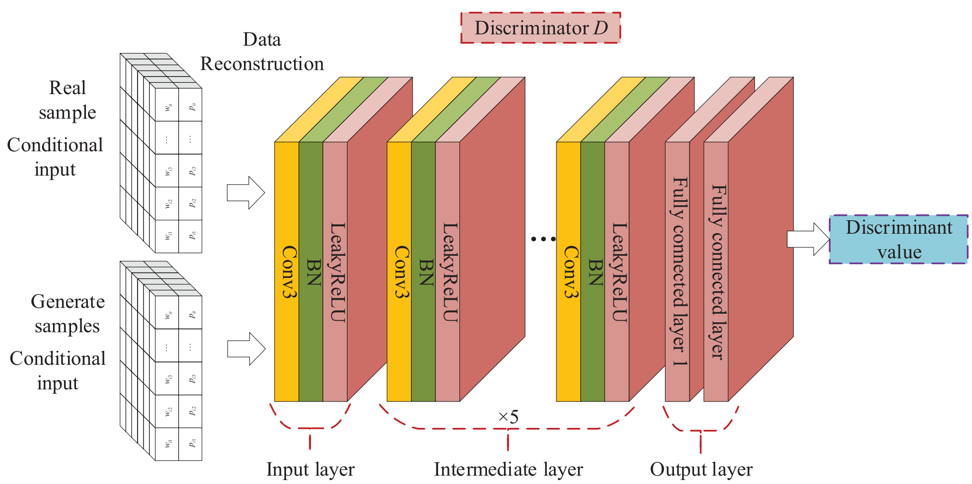

Figure 6: Discriminator network structure

The discriminator is designed with a seven-layer network structure, which is similar to that of the generator network. The first six layers have the structure $Conv3 - BN - LeakyReLU$. The structure of the output layer is $FC_1 - FC_2$, where FC_1 is the first fully connected layer and FC_2 is the second fully connected layer. The convolutional layers all use $3 \times 3 \times 3$ convolution kernels (with a stride of 1 and padding of 1), and the number of channels is 64, 128, 256, 256, 128, and 64 in sequence. The activation function used is $LeakyReLU$.

The architectural design employs small $3 \times 3 \times 3$ convolutional kernels to strengthen the model's capacity for capturing local spatiotemporal patterns. This design principle is widely adopted in deep learning to achieve an optimal balance between receptive field size and computational efficiency [9]. Furthermore, the channel configuration—first increasing then decreasing—facilitates effective feature compression and subsequent reconstruction. This encoder-decoder-like architecture has proven particularly effective for capturing intrinsic local correlations in data [20,21], such as the physical network law-governed power time-series data.

Consequently, the core functionality of the integrated 3D-CNN+CGAN model lies in generating accurate complementary data by leveraging spatiotemporal features extracted from neighboring nodes' conditions.

Considering the possibility of negative values during network training, the generator network consists of 8 layers, with LeakyReLU activation functions used for both the input and middle layers. The output of the $l \in \{1, \cdots, 7\}$ layer of the generator network at time t is shown in Eq. (1).

$$Q_t^l = \text{LeakyReLU}\left[N_{BN}\left(Q_t^{l-1} * f_t^l\right)\right] \tag{1}$$

where Q_t^{l-1} represents the output of the l-1 layer, f_t^l is the 3D convolution kernel representing the l-th layer, and N_{BN} is a batch normalization function.

The discriminator network includes convolutional layers and fully connected layers, and the design of its convolutional layers is like that of the generator network. The output of the $m \in \{1, \cdots, 6\}$ layer of the discriminator network at time t is shown in Eq. (2).

$$\hat{Q}_t^m = \text{LeakyReLU}\left[N_{BN}\left(\hat{Q}_t^{m-1} * \hat{f}_t^m\right)\right] \tag{2}$$

where $\hat{Q}_t^{m-1}$ represents the output of the l-1 layer and $\hat{f}_t^m$ is the 3D convolution kernel representing the m-th layer.

The integration of 3D-Convolutional Neural Networks with Conditional Generative Adversarial Networks (3D-CNN+CGAN) presents a novel and potent solution for time-series recovery in smart grids, advancing beyond established CNN-GAN and Wasserstein GAN frameworks. Conventional 2D-CNN-based GANs often inadequately capture the complex spatio-temporal dependencies inherent in multivariate smart grid data. The key innovation of the 3D-CNN component lies in its direct modeling of data cubes, simultaneously learning correlations across variables, time steps, and contextual dimensions. Concurrently, the CGAN framework enables targeted data generation conditioned on specific operational states (e.g., load profiles), ensuring contextually relevant imputation. This synergy allows the model not merely to fill data gaps with plausible values, but to reconstruct missing segments that preserve the intricate, dynamic integrity of the original system—a critical advancement for reliable data-driven management of smart grid infrastructure.

2.5 Training the 3D-CNN+CGAN Network

During the training process of the 3D-CNN+CGAN network, the weight matrix updates of feature mapping in Fig. 5 and discriminant analysis in Fig. 6 are alternately executed while the generator and discriminator networks are trained, as shown in Fig. 7.

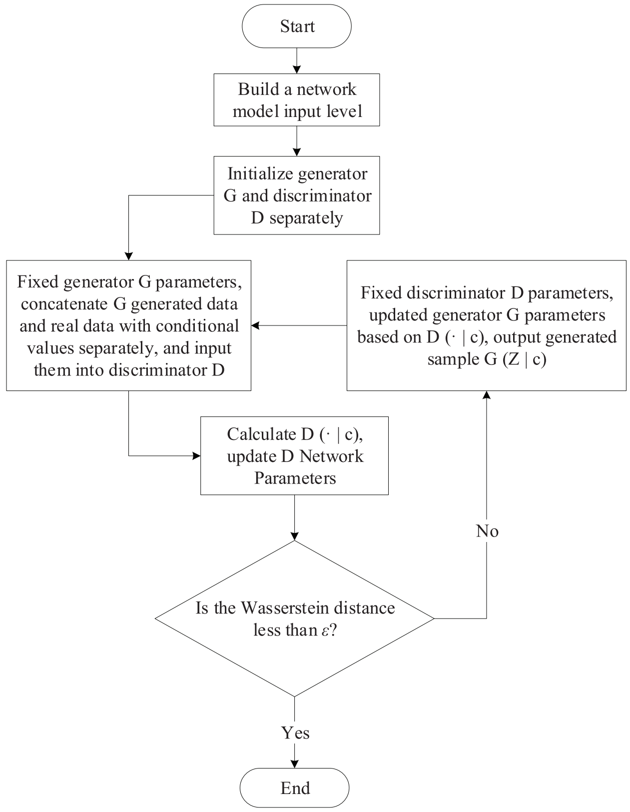

Figure 7: Network training flowchart

In training, the objective of the optimization of the discriminator D is to maximize the difference between the completion value x' and the true value x, with the optimization direction being max $(V(x, x'))$. The goal of the generator G is to minimize the difference between the completion value x' and the true value x, with the optimization direction being min $(V(x, x'))$. Therefore, the optimization objective function of the generative adversarial network is shown in Eq. (3).

$$\min_{G} \max_{D} V(x, x') = E_{x \sim p_{r(x)}} [D(x|c)] - E_{x' \sim p_{(x')}} [D(x'|c)] \tag{3}$$

where x represents the true value, x' represents the generator G completion value, c represents the voltage value of the generator, $p(x \mid c)$ represents the value of x under the given Condition c, $p_r(x)$ represents the true value distribution, $p(x')$ represents the distribution of completion values, $D(\cdot)$ represents the discriminator function, and $V(x, x')$ represents the difference between the completion value x' and the true value x.

To avoid gradient vanishing, the Wasserstein distance is introduced in the discriminator loss function instead of the JS distance to measure the distance between the completion value and the true value. Its definition is shown in Eq. (4).

$$W(p_r(x), p(x')) = \inf_{\gamma \sim \prod(p_r(x), p(x'))} E_{(x, x') \sim \gamma} [\|x - x'\|] \tag{4}$$

where γ represents the joint distribution of an optimal route, W represents the distance between $p_r(x)$ and $p(x')$, $\prod(p_r(x), p(x'))$ represents the joint distribution of $p_r(x)$ and $p(x')$, and $\|x - x'\|$ represents the distance between the true value and the completion value.

3 Dataset Generation

The simulation was conducted on a PC-based computing platform. The core hardware comprised an Intel(R) Core(TM) i7-4720HQ CPU, an NVIDIA Tesla V100 GPU with 32 GB of memory, and 32 GB of DDR5 1600 MHz RAM. All models were developed based on the TensorFlow framework, with training accelerated using CUDA technology.

3.1 Dataset Preparation

As shown in Figs. 8 and 9, 30,000 data sets of time series arrangement sample data were generated using IEEE 14-node and IEEE 33-node typical distribution network test systems, with a sampling frequency of 15 min. Among these, 5% Gaussian noise was added to simulate the presence of distributed power sources in the distribution network; 100 time points were randomly selected, and a 3σ positive deviation was added to simulate the presence of impact loads in the distribution network. For each missing data node, 20,000 data sets from its four neighboring nodes were selected as the training set, and 10,000 data sets were selected as the test set. Based on the empirical values from reference [28], a 10% data loss rate was selected to obtain random/continuously missing data for the test set nodes. Additionally, drawing on the maximum time span of missing data from Reference [29], datasets with missing data rates of 20%,30%, 40% and 50%.

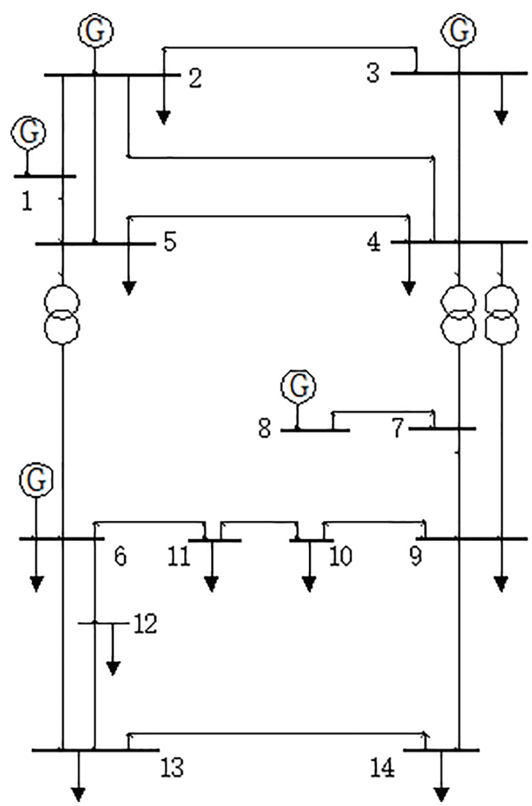

Figure 8: IEEE 14-bus system diagram, where the numbers (1—14) represent the unique identifiers for each bus (node) in the transmission network

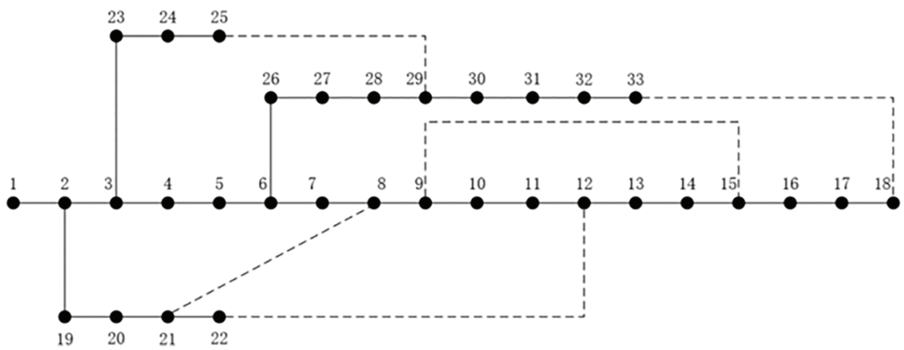

Figure 9: IEEE 33-bus system diagram, where the numbers (1–33) represent the unique identifiers for each bus (node) in the radial distribution network

To clarify applicability, we discuss the correspondence between our simulation settings and real data along three dimensions. (i) **Noise.** PLC environments often show impulsive, heavy-tailed disturbances [30–33]. In our simulations, we use 5% Gaussian perturbations plus occasional 3σ positive deviations to capture background variation and impact-like spikes, keeping SNR within ranges reported in prior work. (ii) **Sampling characteristics.** Although AMI streams can exhibit asynchrony and buffering delays, utilities typically time-align and aggregate data before analysis [34,35]. Our experiments follow this practice with equidistant 15-min sampling (IEEE-14/33), so the model sees uniformly sampled sequences consistent with operational pipelines. (iii) **Data loss.** Field losses are frequently contiguous and often MNAR [33]. We therefore evaluate both pointwise random loss and continuous block loss spanning 10%–50% of each series, approximating the dominant operational dropout patterns.

3.2 Construction of the Data Completion Model

The training set data were trained according to the process shown in Fig. 7 to obtain the generator G and the missing data completion model. The actual application process of the proposed missing data completion method for distribution grids is shown in Fig. 10. Adam optimizer was used to prevent model over fitting with the learning rate set to and the maximum iteration count set to 5000.

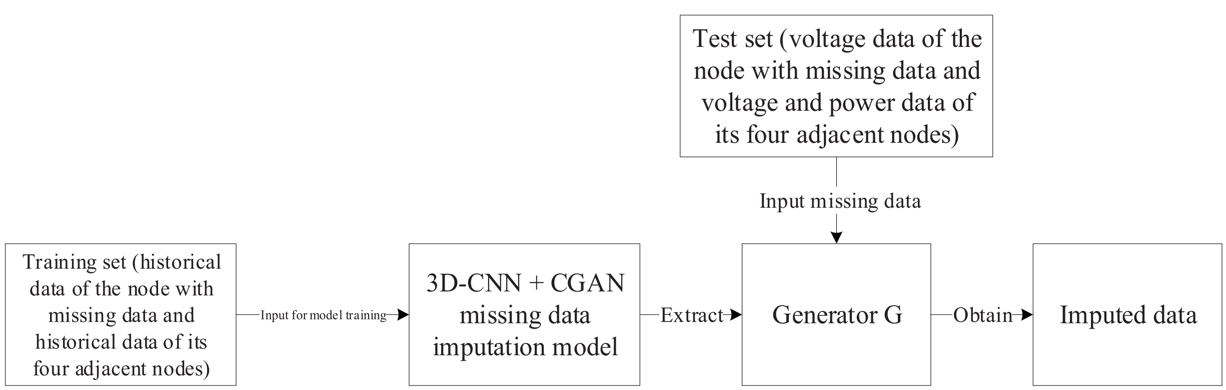

Figure 10: Application flowchart for missing data completion

In this section, the proposed method is compared with five commonly used methods, namely spline interpolation, KNN, RF, GAN, and WGAN, in terms of the root mean square error, spatiotemporal correlation, volatility, and timeliness to demonstrate the data completion quality.

4 Performance Analysis

In this section, we compare the proposed method with the typical methods, such as spline interpolation, KNN, RF, GAN, and WGAN, in terms of the root mean square error, spatiotemporal correlation, volatility, and timeliness, to verify the data completion quality of each testing method.

4.1 Generalizability

The generalizability of a network refers to the effectiveness of a model when faced with unfamiliar data. The completion accuracy of the test data is analyzed by the RMSE [36] in this paper to verify the generalization ability of the constructed model. The calculation of the RMSE is shown in Eq. (5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5}$$

where n represents the amount of data in the test set, i represents the actual time value, $\hat{y}_i$ represents the output value of generator G, and y_i represents the true value of a missing data node.

(1) The average RMSEs of the power completion data with random missing data rates of 10%, 20%, 30%, 40%, and 50% are calculated as shown in Figs. 11 and 12.

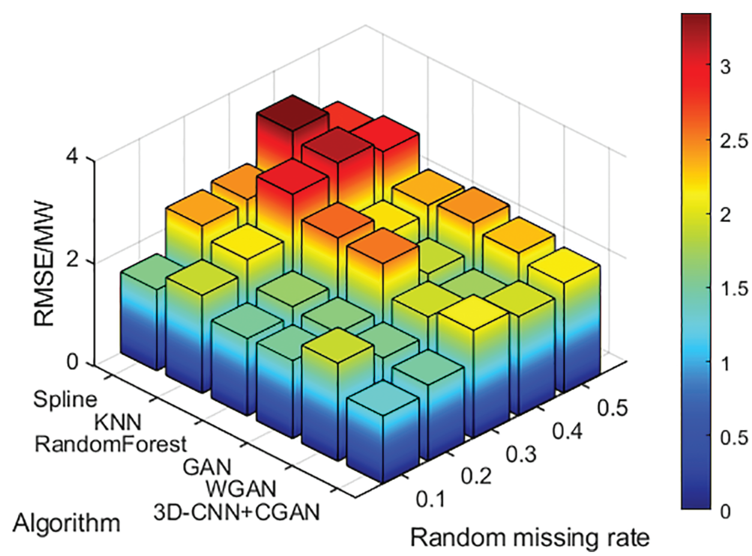

Figure 11: Average RMSEs of different algorithms with random missing data rate (14-bus)

When the random missing data rates are 10%, 20%, 30%, 40%, and 50%, the RMSEs of the tested algorithms increase with increasing missing data rate. Among them, the KNN algorithm has the highest RMSE for completing data, and the 3D-CNN+CGAN algorithm has the lowest RMSE, indicating that the method proposed in this paper has the best generalizability in terms of completing random missing data.

(2) The average RMSEs of the power completion data for continuous missing data rates of 10%, 20%, 30%, 40%, and 50% are calculated as shown in Figs. 13 and 14.

The results in Fig. 12 show that when the continuous missing data rates are 10%, 20%, 30%, 40%, and 50%, the RMSE values of the tested algorithms increase with increasing missing data rate and are greater than the corresponding random missing data rates. The KNN algorithm has the highest RMSE for complete data, whereas the 3D-CNN+CGAN has the lowest RMSE. The RMSE value of the 3D-CNN+CGAN algorithm changes smoothly with increasing missing data rate, indicating that it has strong stability and better generalizability when continuous missing data are complete.

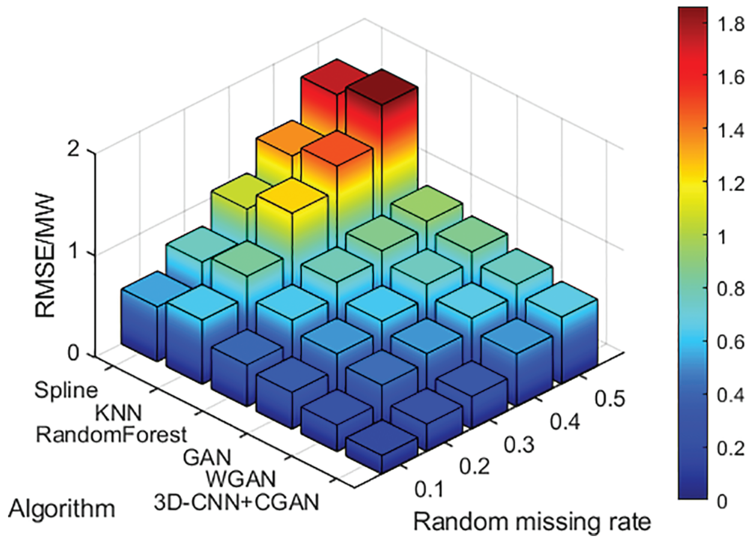

Figure 12: Average RMSEs of different algorithms with random missing data rate (33-bus)

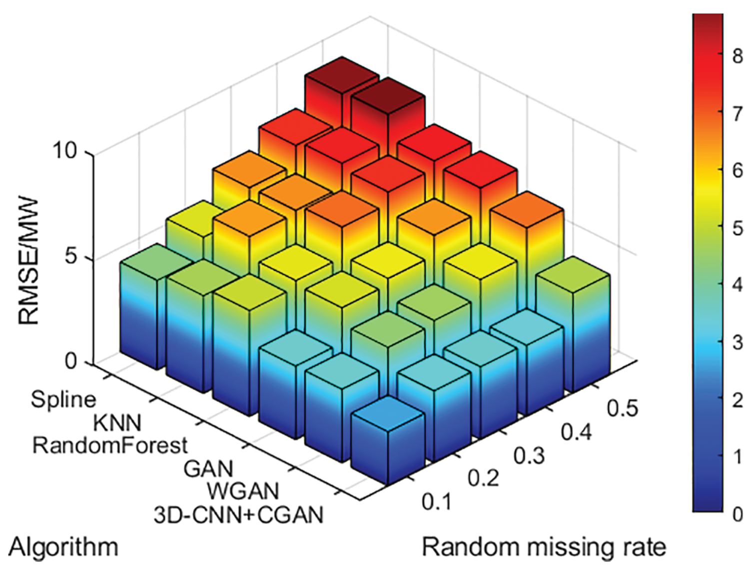

Figure 13: Average RMSEs of different algorithms with continuous missing data rate (14-bus)

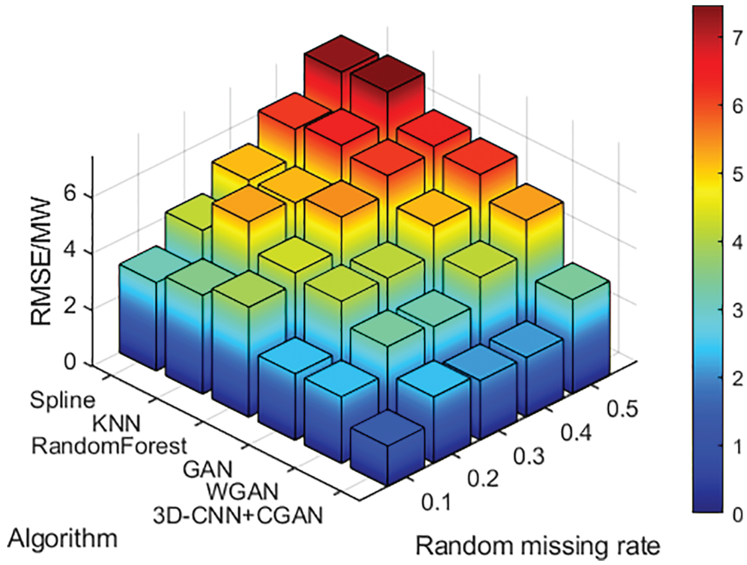

Figure 14: Average RMSEs of different algorithms with continuous missing data rate (33-bus)

4.2 Time Correlation

The autocorrelation function (ACF) value of the actual node power gradually decreases with increasing lag time. Therefore, this article considers the ACF as one of the quality indicators for evaluating the data completeness of the proposed method [37,38]. The expression of the ACF is as Eq. (6):

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k}\left(P_t - \overline{P}\right)\left(P_{t+k} - \overline{P}\right)}{\sum_{t=1}^{n}\left(P_t - \overline{P}\right)^2} \tag{6}$$

where P_t represents the power value at time t, k represents the number of time intervals, $\overline{P}$ represents the mean power, and $\hat{\rho}_k$ represents the power autocorrelation function with a time delay of k.

(1) The ACF of the completion power data is calculated with a random missing data rate of 10%, as shown in Figs. 15 and 16.

The ACF of the 3D-CNN+CGAN algorithm for imputing randomly missing data is closer to the true value and has fewer outliers. The ACF distributions of the WGAN, GAN, RF, and Spline algorithms for imputed data are more dispersed and have more outliers. Although the ACF of the KNN algorithm for imputed data is more concentrated, its median deviates from the true value. This finding indicates that the method proposed in this paper is more stable and reliable for the temporal completion of randomly missing power data.

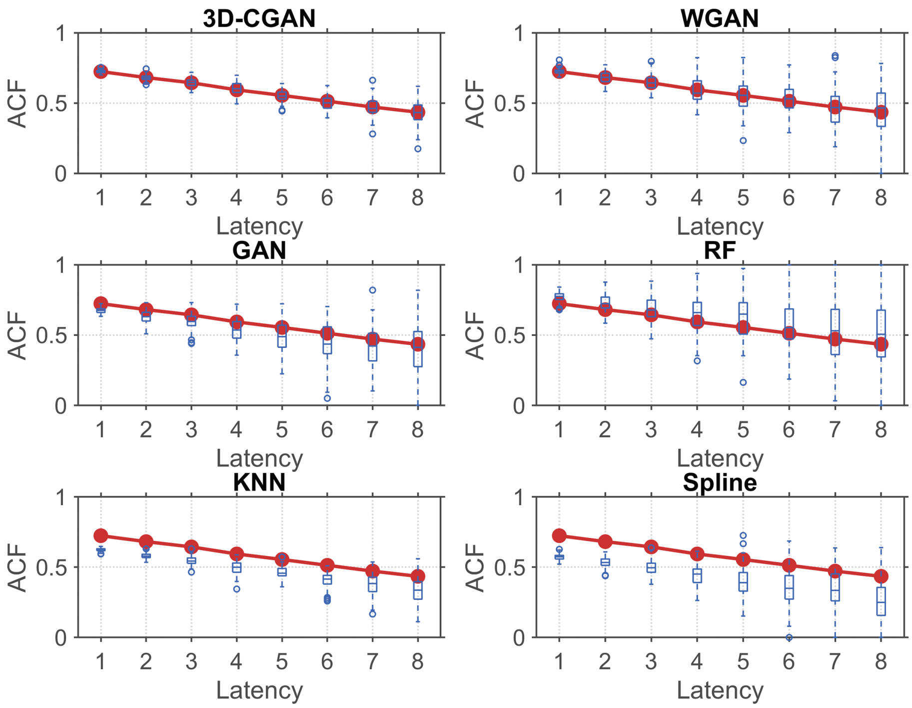

Figure 15: ACF with a 10% random missing data rate (14-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

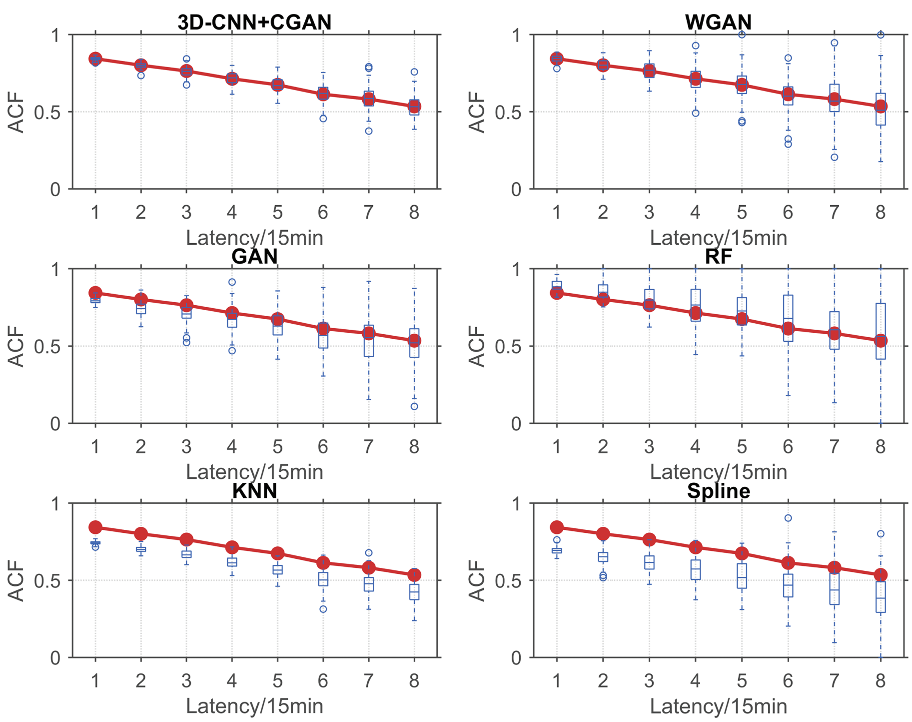

Figure 16: ACF with a 10% random missing data rate (33-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

(2) The ACF of the completion power data is calculated with a continuous missing data rate of 10%, as shown in Figs. 17 and 18.

The median of the ACF completed by the tested algorithm for continuously missing data deviates from the true value, and the ACF distribution completed by the Spline and KNN algorithms can no longer contain the true value. The ACF distributions of the GAN and RF algorithms for completing data are relatively scattered, with many outliers. The ACF distributions of the WGAN and 3D-CNN+CGAN algorithms for completing data can contain most of the true values. The ACF distribution is relatively concentrated, indicating that the method proposed in this paper has higher temporal completion accuracy for continuous missing power data.

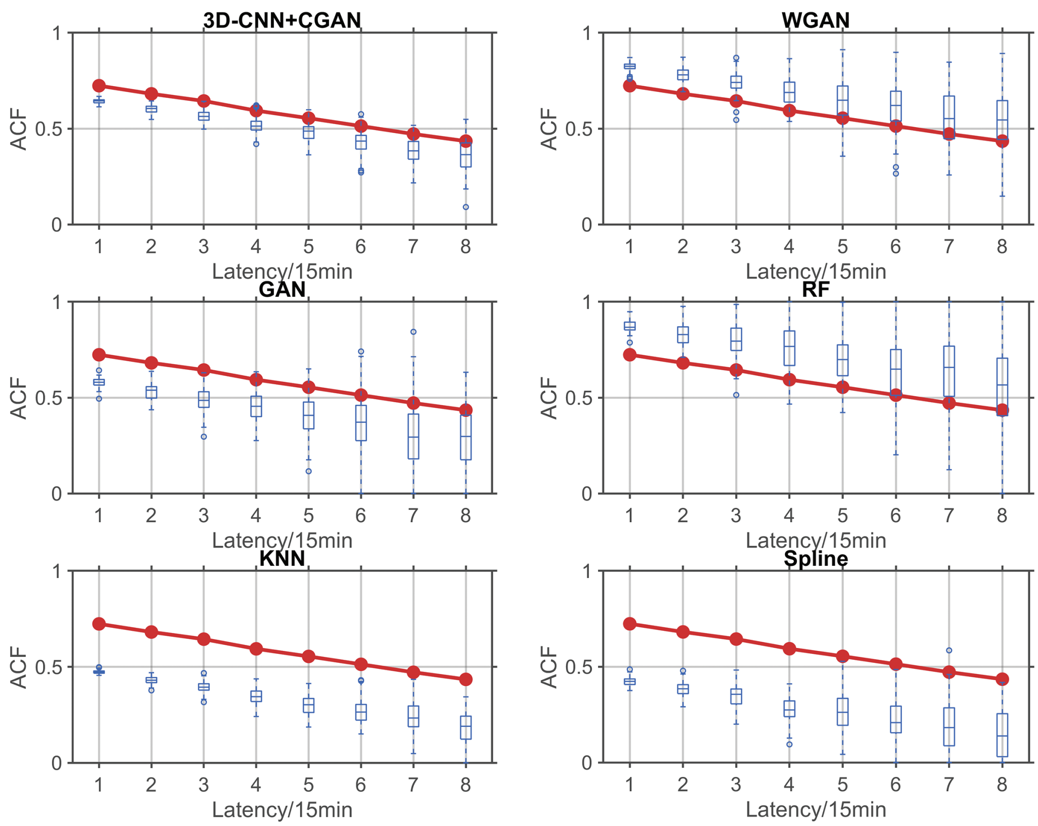

Figure 17: ACF with a 10% continuous missing data rate (14-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

4.3 Spatial Correlation

The Spearman correlation coefficient is a nonparametric statistical method that can be used to measure the spatial correlation between two variables. The calculation of the Spearman correlation coefficient is shown in Eq. (7).

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}\left(R\left(P_{xi}\right) - \overline{R}\left(P_X\right)\right)\left(R\left(P_{yi}\right) - \overline{R}\left(P_Y\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(R\left(P_{xi}\right) - \overline{R}\left(P_X\right)\right)^2 \sum_{i=1}^{n}\left(R\left(P_{yi}\right) - \overline{R}\left(P_Y\right)\right)^2}} \tag{7}$$

where P_X and P_Y represent the actual power of the two nodes; P_{xi} and P_{yi} represent the sample values numbered i; $R\left(P_{xi}\right)$ and $R\left(P_{yi}\right)$ represent the corresponding sequence ranks of samples P_{xi} and P_{yi}, respectively; and $\overline{R}\left(P_X\right)$ and $\overline{R}\left(P_Y\right)$ represent the average rank of sequences P_X and P_Y, respectively.

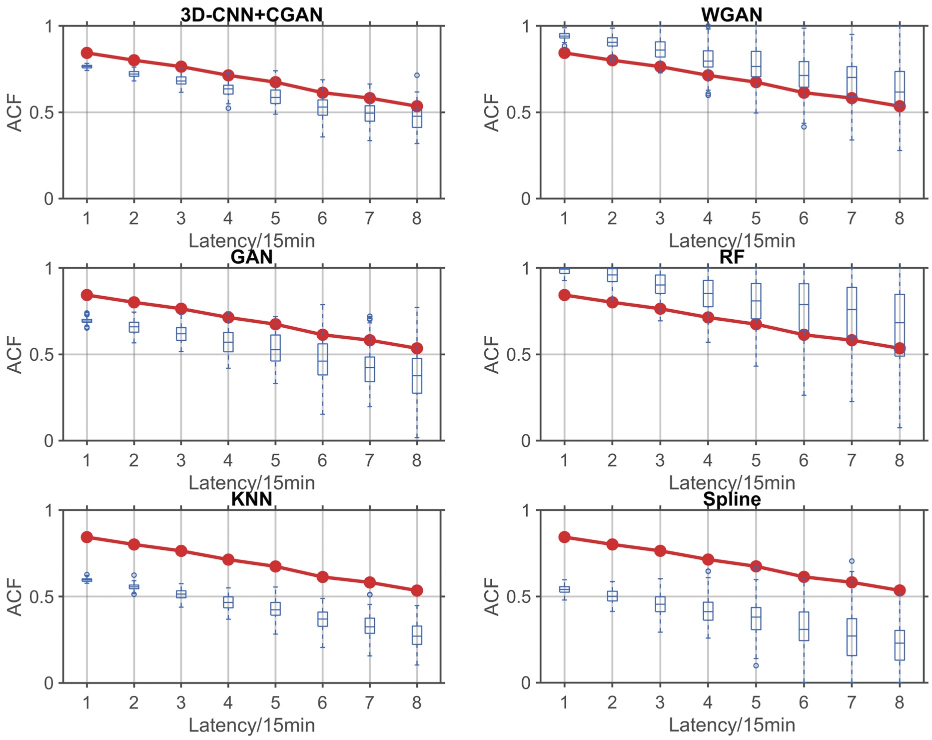

Figure 18: ACF with a 10% continuous missing data rate (33-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

(1) The Spearman correlation coefficient of the completion power data with a random missing data rate of 10% is shown in Figs. 19 and 20.

The Spearman correlation coefficients of the KNN and Spline algorithms for completing randomly missing data deviate significantly from the true values. The GAN algorithm results in fewer outliers in the Spearman correlation coefficient for complete data. The median of the 3D-CNN+CGAN plot is closer to the true value, indicating that the method proposed in this paper has higher spatial correlation completion accuracy for randomly missing power data.

(2) The Spearman correlation coefficient of node completion power data when the continuous missing data rate is 10% (continuous data missing for more than ten days) is calculated as shown in Figs. 21 and 22.

The Spearman correlation coefficients of the RF, KNN, Spline, and GAN algorithms for completing continuously missing data deviate significantly from the true value. The Spearman correlation coefficients between some node combinations are less than 0, indicating that there is no spatial correlation between the completed data and the true value. There are many outliers in the Spearman correlation coefficient of the WGAN algorithm for the complete data. The median of the 3D-CNN+CGAN completion data box plot is closer to the true value, but some overfitting results in a Spearman correlation coefficient of 1, indicating that the performance of the spatial correlation completion with the tested algorithm is worse for the missing power data for more than ten consecutive days.

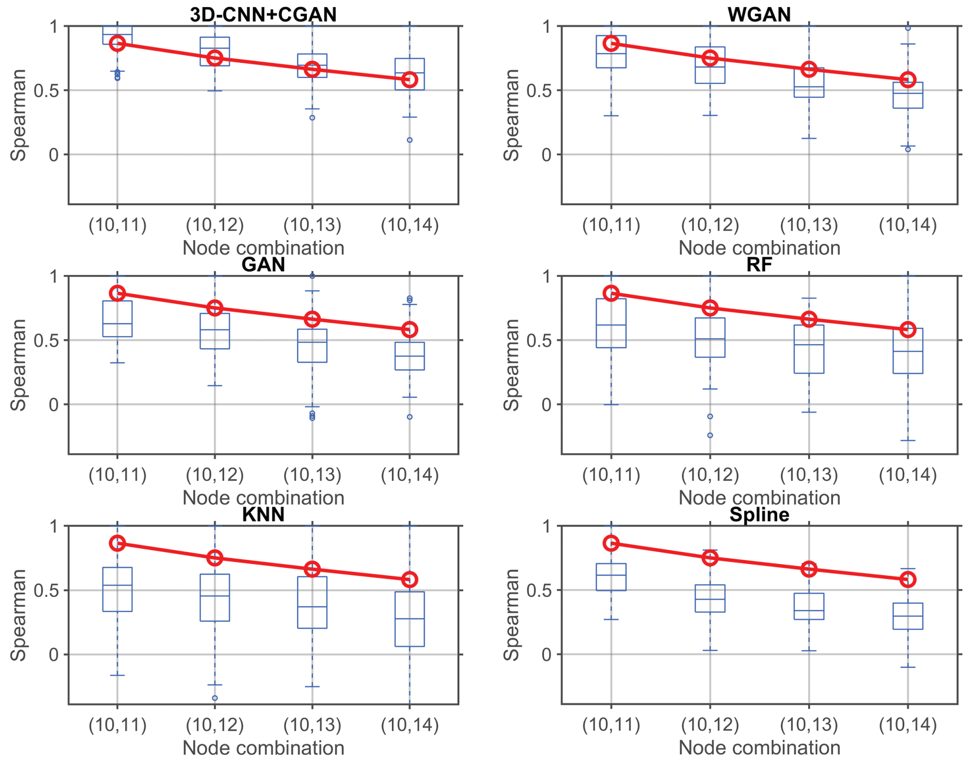

Figure 19: Spearman correlation with a 10% random missing data rate (14-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

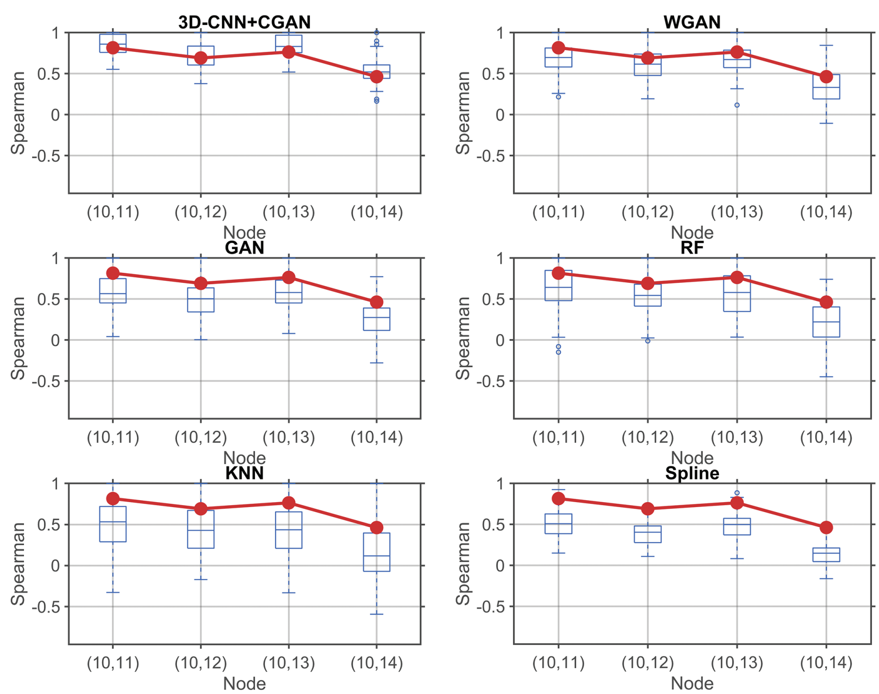

Figure 20: Spearman correlation with a 10% random missing data rate (33-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

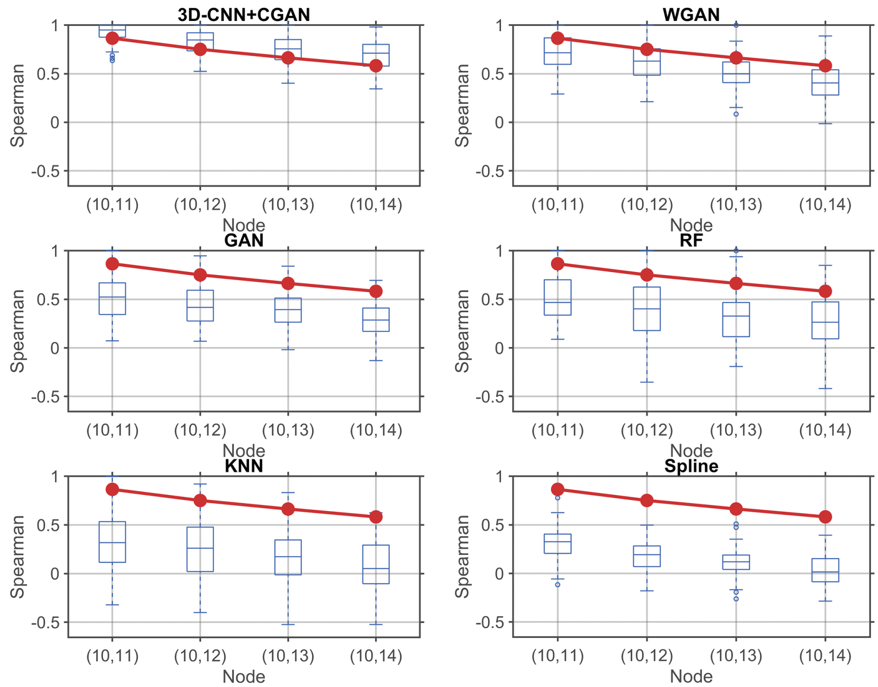

Figure 21: Spearman correlation with a 10% continuous missing data rate (14-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

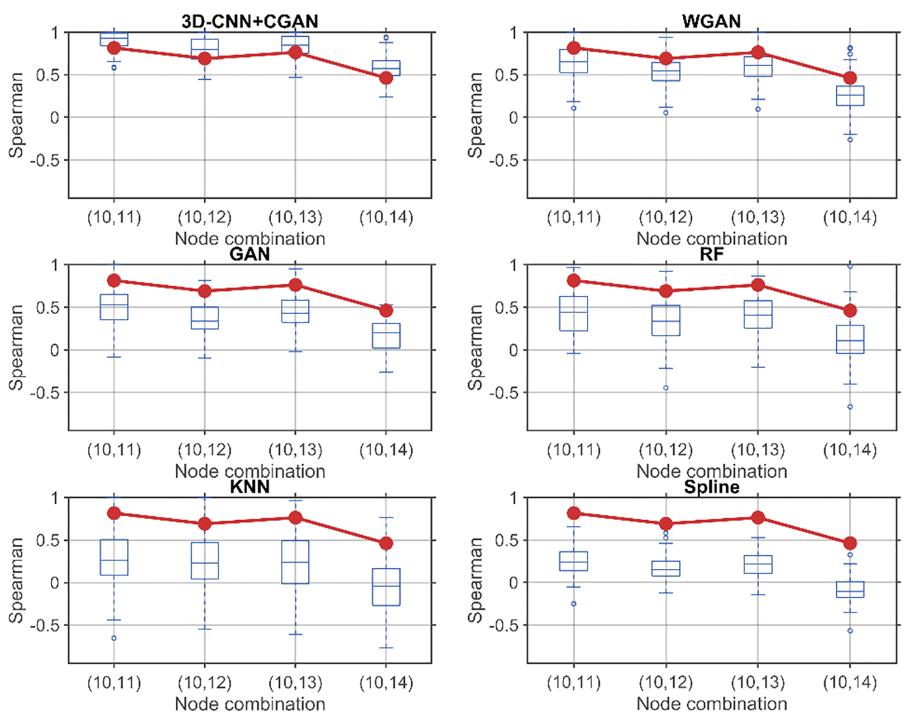

Figure 22: Spearman correlation with a 10% continuous missing data rate (33-bus) (nodes in red (⊖) represent true node, and nodes in blue (◯) represent abnormal node)

4.4 Volatility

The maximum range (MR) (denoted as R_{MR}) refers to the difference between the maximum and minimum values within time T and is defined and calculated as shown in Eq. (8).

$$R_{MR} = \begin{cases} \dfrac{P_{t\max} - P_{t\min}}{P_N}, t_{P_{t\max}} > t_{P_{t\min}} \\ \dfrac{P_{t\min} - P_{t\max}}{P_N}, t_{P_{t\max}} < t_{P_{t\min}} \end{cases} \tag{8}$$

where $P_{t\max}$ and $P_{t\min}$ represent the maximum and minimum values of node power, respectively, during time T; $t_{P_{t\max}}$ and $t_{P_{t\min}}$ represent the time length from the initial time of the time series of $P_{t\max}$ and $P_{t\min}$, respectively; and P_N represents the network benchmark capacity.

(1) The maximum fluctuation range of the completion power data when the random missing data rate is 10% is calculated as shown in Figs. 23 and 24.

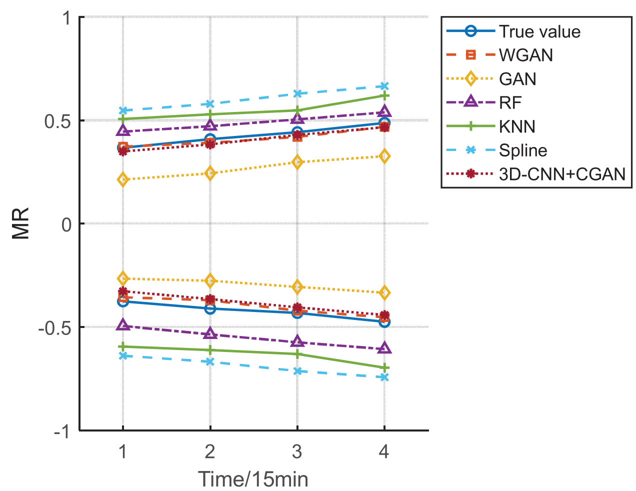

Figure 23: MR variation range among different algorithms with a 10% random missing data rate (14-bus)

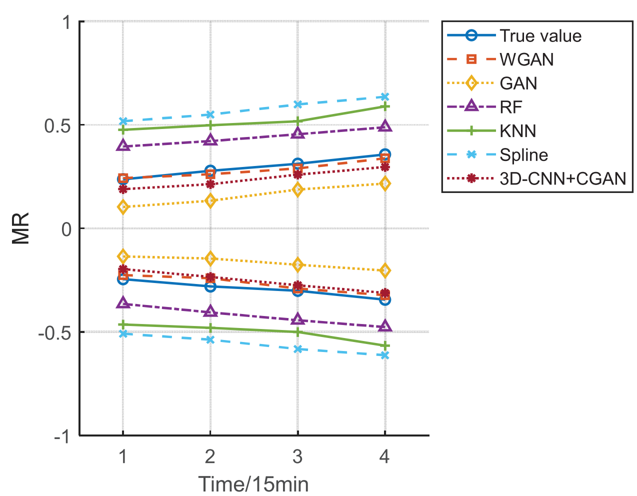

Figure 24: MR variation range among different algorithms with a 10% random missing data rate (33-bus)

At different times T, the maximum fluctuation amplitude of the Spline, KNN, and RF algorithms for completing randomly missing data is significantly greater than the true value. The maximum fluctuation amplitude of the GAN algorithm for completing random missing data is slightly smaller than the true value. The maximum fluctuation amplitude range of the WGAN and 3D-CNN+CGAN algorithms is close to the true value. The method proposed in this article is more in line with the true fluctuation characteristics of node power in completing randomly missing power data.

(2) The maximum fluctuation range of the completion power data when the continuous missing data rate is 10% is calculated as shown in Figs. 25 and 26.

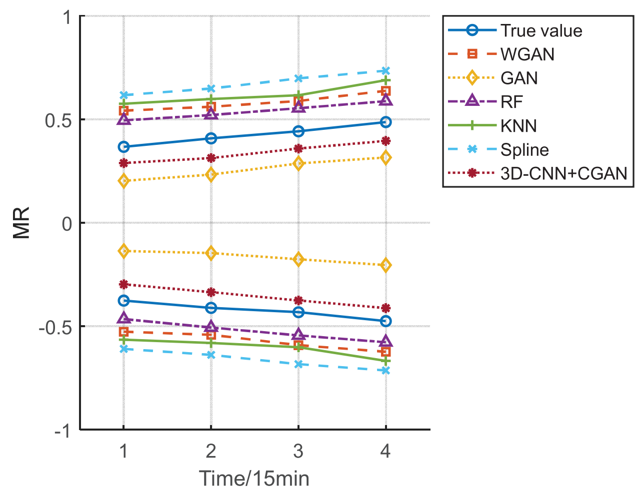

Figure 25: MR variation range among different algorithms with a 10% continuous missing data rate (14-bus)

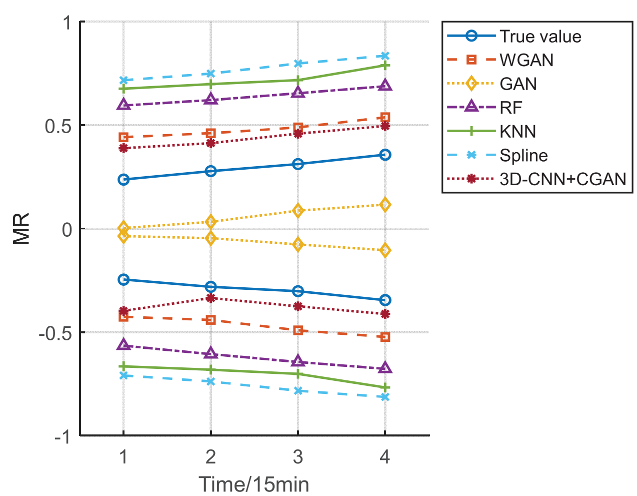

Figure 26: MR variation range among different algorithms with a 10% continuous missing data rate (33-bus)

In Fig. 26, compared with that of the GAN algorithm, the maximum fluctuation amplitude with the completion of continuous missing data of the Spline, KNN, RF, WGAN, and 3D-CNN+CGAN algorithms is much greater at different times T. The maximum fluctuation amplitude range of the 3D-CNN+CGAN algorithm is closer to the true value, indicating that the method proposed in this paper is more in line with the fluctuation characteristics of actual node power for the completion of continuous missing power data and is more precise.

4.5 Timeliness

To test the feasibility of the method proposed in this article, 3D-CNN+CGAN was compared with several commonly used methods in terms of timeliness, and the results are shown in Table 1.

Table 1: Time effectiveness analysis of data estimation algorithms

Data Completion Algorithm	Time-Consuming/s	
	IEEE 14-Node	IEEE 33-Node
Spline	0	0
KNN	513.64	526.36
RF	559.16	572.49
GAN	723.64	741.36
WGAN	623.64	653.64
3D-CNN+CGAN	603.49	630.16

The 3D-CNN+CGAN model requires slightly longer training time per iteration than WGAN due to its complex architecture. However, by incorporating Wasserstein distance and conditional voltage information, it achieves faster convergence and more stable performance compared to both GAN and WGAN.

By comprehensively analyzing the five indicators of generalizability, temporal correlation, spatial correlation, volatility, and timeliness, the comprehensive comparison results of data completion quality between the proposed method and the testing methods are obtained, as shown in Fig. 27. The numbers 1–6 are used to indicate the degree of superiority or inferiority of a certain indicator for each method, with 1 representing the worst and 6 representing the best.

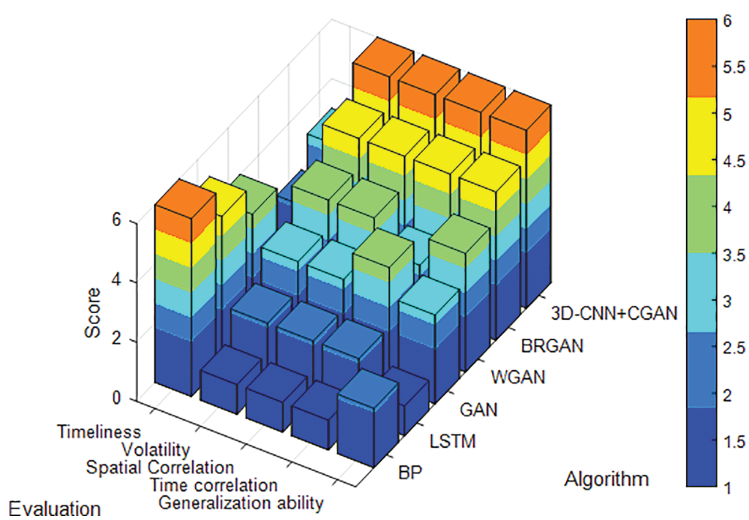

Figure 27: Comparison of the comprehensive performance of various algorithms

With respect to the random and continuous missing node power data, the spline interpolation, KNN and random forest algorithms have fast data completion speeds, but the completion quality is poor. Owing to the ability of the random forest algorithm to construct a prediction model based on known node voltage data to complete missing power data, the data completion quality is better than that of the other two algorithms; the GAN, WGAN and 3D-CNN+CGAN have high data completion quality because of the spatiotemporal correlation between adjacent node power data fusion. The WGAN and 3D-CNN+CGAN algorithms improve the stability of the training process and reduce the vanishing gradient problem by introducing the Wasserstein distance to optimize the objective function. They outperform the GAN in terms of complex spatial relationships and timeliness between training and learning nodes. Additionally, the model utilizes node voltage data to guide training, ensuring that the reconstructed data satisfies distributional mapping constraints under specific conditions, thereby improving accuracy in handling continuous missing data. The 3D-CNN architecture's ability to extract spatiotemporal features and leverage conditional voltage information accelerates model convergence through enhanced machine learning support.

5 Conclusions

With the extensive integration of distributed renewable energy and energy storage systems into distribution networks, the volume of situational awareness data has increased dramatically. To address the issue of random and continuous data gaps caused by limitations in communication bandwidth or failures in transmission and storage, in this paper, a data completion method was proposed based on multidimensional feature fusion and data augmentation, where a 3D-CNN was employed to extract the spatiotemporal features from measurement data and a CGAN was used to model the complex mapping relationships. The main findings are summarized as follows:

(1) Compared with baseline algorithms, the proposed method achieves lower root mean square errors (RMSEs) across various missing data rates in the completion of power data for distribution network nodes. In the scenarios involving continuous missing data, the RMSE of the completed data increases gradually with increasing missing data rate, demonstrating the robustness and reliability of the method.
(2) The spatiotemporal correlations between missing nodes and their neighboring measurements were captured more accurately than benchmark algorithms, leading to improved performance in terms of the completion of both random and continuous missing data.
(3) The standard GAN method significantly underestimates the fluctuation amplitude of node data, as do methods such as Spline, KNN, and RF. The WGAN and the proposed method generate complete data that more closely align with the true fluctuation characteristics.

The integration of 3D-CNN +CGAN enhances feature representation, but it increases the complexity of the model. Our future work will be involved in optimizing the model structure for improved computational efficiency, validating its performance on larger-scale real-world datasets, exploring attention-based network architectures to enhance key feature capture, and so on.

Acknowledgement: I would like to express my heartfelt gratitude to the other authors for their invaluable contributions and assistance in this research work.

Funding Statement: This research was funded by the Science and Technology Project of China Southern Power Grid Co., Ltd. (Grant no. GXKJXM20222165).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Kewen Li and Xiaoyong Yu; methodology, Kewen Li; software, Kewen Li; validation, Kewen Li, Xiaoyong Yu and Shifeng Ou; formal analysis, Kewen Li; investigation, Kewen Li; resources, Kewen Li; data curation, Kewen Li and Jueming Pan; writing—original draft preparation, Kewen Li; writing—review and editing, Kewen Li; visualization, Kewen Li; supervision, Kewen Li and Jueming Pan; project administration, Kewen Li; funding acquisition, Kewen Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Liang G, Liu G, Zhao J, Liu Y, Gu J, Sun G, et al. Super resolution perception for improving data completeness in smart grid state estimation. Engineering. 2020;6(7):789–800. doi:10.1016/j.eng.2020.06.006.
2. Mohammadrezaee R, Ghaisari J, Yousefi G, Kamali M. Dynamic state estimation of smart distribution grids using compressed measurements. IEEE Trans Smart Grid. 2021;12(5):4535–42. doi:10.1109/tsg.2021.3071514.
3. Zhang D, Jin X, Shi P, Chew X. Real-time load forecasting model for the smart grid using Bayesian optimized CNN-BiLSTM. Front Energy Res. 2023;11:1193662. doi:10.3389/fenrg.2023.1193662.
4. Şen M, Üstün Ercan S. The use of data transmission technique via power line communication. Balkan J Electr Comput Eng. 2024;12(2):144–51. doi:10.17694/bajece.1432481.
5. Honda K, Ichihashi H. Linear fuzzy clustering techniques with missing values and their application to local principal component analysis. IEEE Trans Fuzzy Syst. 2004;12(2):183–93. doi:10.1109/tfuzz.2004.825073.
6. Kim M, Park S, Lee J, Joo Y, Choi J. Learning-based adaptive imputation method with kNN algorithm for missing power data. Energies. 2017;10(10):1668. doi:10.3390/en10101668.
7. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min. 2017;10(6):363–77. doi:10.1002/sam.11348.
8. Pisner DA, Schnyer DM. Support vector machine. In: Machine learning. Cambridge, MA, USA: Academic Press; 2020. p. 101–21. doi:10.1016/b978-0-12-815739-8.00006-7.
9. Du W, Côté D, Liu Y. SAITS: self-attention-based imputation for time series. Expert Syst Appl. 2023;219(10):119619. doi:10.1016/j.eswa.2023.119619.
10. Cao W, Wang D, Li J, Zhou H, Li L, Li Y. BRITS: bidirectional recurrent imputation for time series. arXiv:1805.10572. 2018. doi:10.48550/arxiv.1805.10572.
11. Tashiro Y, Song J, Song Y, Ermon S. CSDI: conditional score-based diffusion models for probabilistic time series imputation. arXiv:2107.03502. 2021. doi:10.48550/arxiv.2107.03502.
12. Cao YJ, Jia LL, Chen YX, Lin N, Yang C, Zhang B, et al. Recent advances of generative adversarial networks in computer vision. IEEE Access. 2019;7:14985–5006. doi:10.1109/access.2018.2886814.
13. Jia J, Xu T, Li M, Li X, Pan W. Data recovery of power plant platform based on generative adversarial network. In: Proceedings of the 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA); 2023 Feb 24–26; Changchun, China. New York, NY, USA: IEEE; 2023. doi:10.1109/eebda56825.2023.10090824.
14. Tun TP, Pisica I. Synthetic electricity consumption data generation using tabular generative adversarial networks. In: Proceedings of the 2023 58th International Universities Power Engineering Conference (UPEC); 2023 Aug 30–Sep 1; Dublin, Ireland. New York, NY, USA: IEEE; 2023. doi:10.1109/upec57427.2023.10294666.
15. Luo YH, Cai XR, Zhang Y, Xu J, Yuan X. Multivariate time series imputation with generative adversarial networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montreal, QC, Canada. Vancouver, BC, Canada: Curran Associates Inc.; 2018. p. 1603–14.
16. Khaled A, Elsir AMT, Shen Y. TFGAN: traffic forecasting using generative adversarial network with multi-graph convolutional network. Knowl Based Syst. 2022;249:108990. doi:10.1016/j.knosys.2022.108990.

17. Ren C, Xu Y. A fully data-driven method based on generative adversarial networks for power system dynamic security assessment with missing data. IEEE Trans Power Syst. 2019;34(6):5044–52. doi:10.1109/tpwrs.2019.2922671.
18. Yan X, Huang D, Zheng Z, Zheng W, Chen P. Load forecasting method for electric vehicles based on CNN-GRU-attention model. In: Proceedings of the 2024 IEEE 5th International Conference on Advanced Electrical and Energy Systems (AEES); 2024 Nov 29–Dec 1; Lanzhou, China. New York, NY, USA: IEEE; 2024. doi:10.1109/aees63781.2024.10872525.
19. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784. 2014. doi:10.48550/arxiv.1411.1784.
20. Yoon J, Jordon J, Van der Schaar M. GAIN: missing data imputation using generative adversarial nets. arXiv:1806.02920. 2018. doi:10.48550/arxiv.1806.02920.
21. Fang J, Zheng L, Liu C. A novel method for missing data reconstruction in smart grid using generative adversarial networks. IEEE Trans Ind Inf. 2024;20(3):4408–17. doi:10.1109/tii.2023.3306366.
22. Liu X, Zhang Z. A two-stage deep autoencoder-based missing data imputation method for wind farm SCADA data. IEEE Sens J. 2021;21(9):10933–45. doi:10.1109/jsen.2021.3061109.
23. Sajjad M, Khan ZA, Ullah A, Hussain T, Ullah W, Lee MY, et al. A novel CNN-GRU-based hybrid approach for short-term residential load forecasting. IEEE Access. 2020;8:143759–68. doi:10.1109/access.2020.3009537.
24. Zhang Y, Tennakoon T, Chan YH, Chan KC, Fu SC, Tso CY, et al. Energy consumption modelling of a passive hybrid system for office buildings in different climates. Energy. 2022;239(12):121914. doi:10.1016/j.energy.2021.121914.
25. Almutairi K, Hosseini Dehshiri SS, Hosseini Dehshiri SJ, Mostafaeipour A, Jahangiri M, Techato K. Technical, economic, carbon footprint assessment, and prioritizing stations for hydrogen production using wind energy: a case study. Energy Strategy Rev. 2021;36(58):100684. doi:10.1016/j.esr.2021.100684.
26. Yu M, Zhang S, Shi F, Sun J, Wei J, Wu Y, et al. Key transmission section search based on graph theory and PMU data for vulnerable line identification in power system. J Electr Comput Eng. 2023;2023(11):8643537. doi:10.1155/2023/8643537.
27. Wang Y, Shen X, Tang X, Liu J. A joint estimation method of distribution network topology and line parameters based on power flow graph convolutional networks. Energies. 2024;17(21):5272. doi:10.3390/en17215272.
28. González-Muñoz M, Dominguez-Benetton X, Domínguez-Maldonado J, Valdés-Lozano D, Pacheco-Catalán D, Ortega-Morales O, et al. Polarization potential has no effect on maximum current density produced by halotolerant bioanodes. Energies. 2018;11(3):529. doi:10.3390/en11030529.
29. Andreotti A, Araneo R, Pierno A. A survey on analytical solutions and tools for lightning-induced voltages calculations. Electr Power Syst Res. 2021;194(1):107104. doi:10.1016/j.epsr.2021.107104.
30. Zimmermann M, Dostert K. Analysis and modeling of impulsive noise in broad-band powerline communications. IEEE Trans Electromagn Compat. 2002;44(1):249–58. doi:10.1109/15.990732.
31. Katayama M, Yamazato T, Okada H. A mathematical model of noise in narrowband power line communication systems. IEEE J Select Areas Commun. 2006;24(7):1267–76. doi:10.1109/jsac.2006.874408.
32. Galli S, Scaglione A, Wang Z. For the grid and through the grid: the role of power line communications in the smart grid. Proc IEEE. 2011;99(6):998–1027. doi:10.1109/jproc.2011.2109670.
33. Bai T, Zhang H, Wang J, Xu C, Elkashlan M, Nallanathan A, et al. Fifty years of noise modeling and mitigation in power-line communications. IEEE Commun Surv Tutor. 2021;23(1):41–69. doi:10.1109/comst.2020.3033748.
34. Peppanen J, Zhang X, Grijalva S, Reno MJ. Handling bad or missing smart meter data through advanced data imputation. In: Proceedings of the 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT); 2016 Sep 6–9; Minneapolis, MN, USA. New York, NY, USA: IEEE; 2016. doi:10.1109/isgt.2016.7781213.
35. Jeng RS, Kuo CY, Ho YH, Lee MF, Tseng LW, Fu CL, et al. Missing data handling for meter data management system. In: Proceedings of the Fourth International Conference on Future Energy Systems; 2013 May 21–24; Berkeley, CA, USA. New York, NY, USA: ACM; 2013. doi:10.1145/2487166.2487204.
36. Kumar P, Purushothaman KE, Arunmozhi RV, Jayachitra S, Rani DL, Kaliappan S. Fault detection in solar power system with Internet of Things using multi resolution sinusoidal neural network—snow geese optimization

approach. In: Proceedings of the 2024 4th International Conference on Sustainable Expert Systems (ICSES); 2024 Oct 15–17; Kaski, Nepal. New York, NY, USA: IEEE; 2024. doi:10.1109/icses63445.2024.10763318.
37. Sun B, Luo F, Jin X, Yang Z. Power data integrity verification model based on blockchain technology. In: Proceedings of the 2023 International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII); 2023 Jun 9–11; Melbourne, Australia. New York, NY, USA: IEEE; 2023. doi:10.1109/icmiii58949.2023.00075.
38. Huang YC, Lin HC, Huang YW. Application of the autocorrelation function to working-day calculation in power management. Appl Soft Comput. 2018;68(1):923–31. doi:10.1016/j.asoc.2017.11.033.