

OPEN ACCESS

EDITED BY

Manoj Kumar Nallapaneni,
Maastricht University, Netherlands

REVIEWED BY

Kenneth E. Okedu,
Melbourne Institute of Technology, Australia
Javier Tarrío-Saavedra,
Universidade da Coruña, Spain

*CORRESPONDENCE

Wenjie Xue,
✉ 15138656738@163.com

RECEIVED 25 November 2024

REVISED 03 November 2025

ACCEPTED 17 December 2025

PUBLISHED 16 January 2026

CITATION

Xue W, An C, Sheng T, Dong P, Zhai Y and Zhang J (2026) A self-attention enhanced GRU network for predicting life cycle cost of substation GIS equipment.
Front. Energy Res. 13:1533878.
doi: 10.3389/fenrg.2025.1533878

COPYRIGHT

© 2026 Xue, An, Sheng, Dong, Zhai and Zhang. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A self-attention enhanced GRU network for predicting life cycle cost of substation GIS equipment

Wenjie Xue^{1*}, Chaoyin An², Tengfei Sheng², Pingxian Dong¹, Yuxin Zhai¹ and Jinfeng Zhang¹

¹State Grid Henan Electric Power Company Economic Research Institute, Zhengzhou, Henan, China,

²State Grid Henan Electric Power Company, Zhengzhou, Henan, China

The challenges of limited sample size and anomalies in the life cycle cost (LCC) data of substation GIS equipment make it difficult to achieve accurate LCC estimation. To address these issues, this paper introduces similar substation GIS equipment LCC data and incorporates a self-attention mechanism to expand the sample size. Additionally, a GRU algorithm is applied to predict time series data and mitigate the impact of anomalies on the prediction. A novel SA-GRU-based model for substation GIS equipment LCC data prediction is proposed. Considering the discrepancies in data characteristics between similar substation GIS equipment LCC data and existing sample data, the self-attention mechanism is utilized to enhance data features. The influence weights between similar and sample data are quantified based on these enhanced features to ensure consistency in data characteristics. Furthermore, the GRU algorithm is employed to alleviate gradient vanishing and exploding issues during training, ultimately enabling LCC data prediction over the entire life cycle. Validation using LCC data from a 110 kV GIS equipment substation in Henan Province demonstrates that the proposed method significantly improves prediction accuracy compared to other prediction approaches.

KEYWORDS

substation GIS equipment, LCC data, forecasting model, self-attention mechanism, GRU model

1 Introduction

The results of Life Cycle Cost (LCC) estimation directly determine the selection of GIS equipment in substations (Cai et al., 2011; Su et al., 2012). Traditional LCC estimation methods, which rely on manual calculations, have gradually been replaced by predictive methods based on LCC data. However, the accuracy of LCC data prediction depends heavily on the continuous collection of substantial economic data generated by GIS equipment throughout its lifecycle. Currently, the collection of LCC data for GIS equipment in substations in China started relatively late, and the data typically suffers from issues such as anomalies and insufficient sample size. These challenges hinder accurate LCC estimation (Shi et al., 2020; Li et al., 2023). Therefore, developing a prediction method suitable for LCC data of substation GIS equipment is critical for addressing the economic evaluation and selection of GIS equipment under existing conditions (Du, 2013).

In recent years, extensive research has focused on substation LCC prediction. Xiong Z. et al. (2021) developed a substation LCC prediction model by optimizing a least-squares support vector machine (LS-SVM) with quantum-behaved particle swarm

optimization (QPSO), effectively improving accuracy under small-sample conditions. Qiao et al. (2015) used a genetic algorithm (GA) to tune LS-SVM hyperparameters, further enhancing generalization and predictive stability. Addressing the unique environment of high-latitude frigid regions, Liu L. et al. (2020) constructed an LCC prediction model in which LS-SVM is optimized by the salp swarm algorithm (SSA), achieving a marked gain in region-specific accuracy. To handle data uncertainty, Liu S. et al. (2020) incorporated fuzzy theory into the LCC estimation process, improving the reliability of the estimates. Wang et al. (2021) established an LCC mathematical model for a specific substation type and conducted sensitivity analysis to identify key drivers.

Despite these advances in traditional model optimization and uncertainty handling, challenges remain for GIS equipment LCC in substations, a task characterized by small samples and nonstationary time series. Early studies (Wang et al., 2020) based on optimized least squares are susceptible to missing values and outliers, leading to substantial errors (Kong et al., 2020). Although fuzzy neural networks (Yang et al., 2017; Zhu et al., 2024) can mitigate the impact of anomalies, directly applying neural networks to GIS LCC time series often results in parameter instability during training, making vanishing or exploding gradients more likely (Xiong Y. et al., 2021).

In the related domain of wind power forecasting—where data exhibit similar characteristics—research is comparatively mature (Peng et al., 2016; Tao et al., 2018). For instance (Xue et al., 2019), introduced the gated recurrent unit (GRU), whose gating mechanism strengthens temporal dependencies. However, GRU models require sizable training datasets, and their accuracy is difficult to guarantee in small-sample settings (Zhang et al., 2022). To address this (Wang et al., 2022), integrated an attention mechanism with GRU and enlarged the training set using data from similar wind farms. Nevertheless, the method relies on manual feature extraction; when data distributions are complex, feature design becomes challenging and subjective, and accuracy remains insufficient (Yang et al., 2023).

Contributions. This paper makes the following contributions to LCC forecasting for substation GIS equipment:

1. Model design for data scarcity and anomalies. We propose SA-GRU, a self-attention-enhanced GRU tailored to LCC time series with scarce samples and outliers.
2. Attention-weighted similar-station integration. The self-attention module automatically learns feature representations and assigns data-driven weights to existing and similar-station samples, yielding effective sample augmentation and improved robustness.
3. Real-world validation and deployment. We evaluate the approach on real LCC data and report a successful deployment in a technical retrofit and major-overhaul project at a substation in Henan Province, demonstrating practical utility.

For clarity, our forecasting target is the annualized life-cycle cost (LCC), which serves as the decision-level quantity for substation equipment selection and budgeting. As LCC aggregates capital expenditure, routine/corrective maintenance, overhaul, outage-related loss, and end-of-life components, modeling it directly matches operational practice. The remainder of this paper is organized as follows. Section 2 describes the dataset and formalizes the problem; Section 3 details the SA-GRU methodology; Section 4

reports the experimental setup and comparative results; Section 5 concludes and discusses future work.

2 LCC data of GIS equipment in substations and data prediction methods

This chapter establishes the foundational groundwork for GIS equipment LCC prediction by addressing two key prerequisites: first, analyzing the inherent characteristics of LCC data to support the design of effective methods; second, examining the limitations of current data prediction techniques and suggesting targeted improvements.

2.1 LCC data of GIS equipment in substations

The economic performance of GIS equipment is typically assessed based on its Life Cycle Cost (LCC). As shown in Figure 1, the LCC of GIS equipment consists of several types of costs: initial investment cost, operation cost, maintenance cost, failure cost, and decommissioning cost. Among these, the initial investment cost and decommissioning cost can be largely determined when the substation is completed, making them fixed costs. However, the operation, maintenance, and failure costs require annual statistics and are considered variable costs.

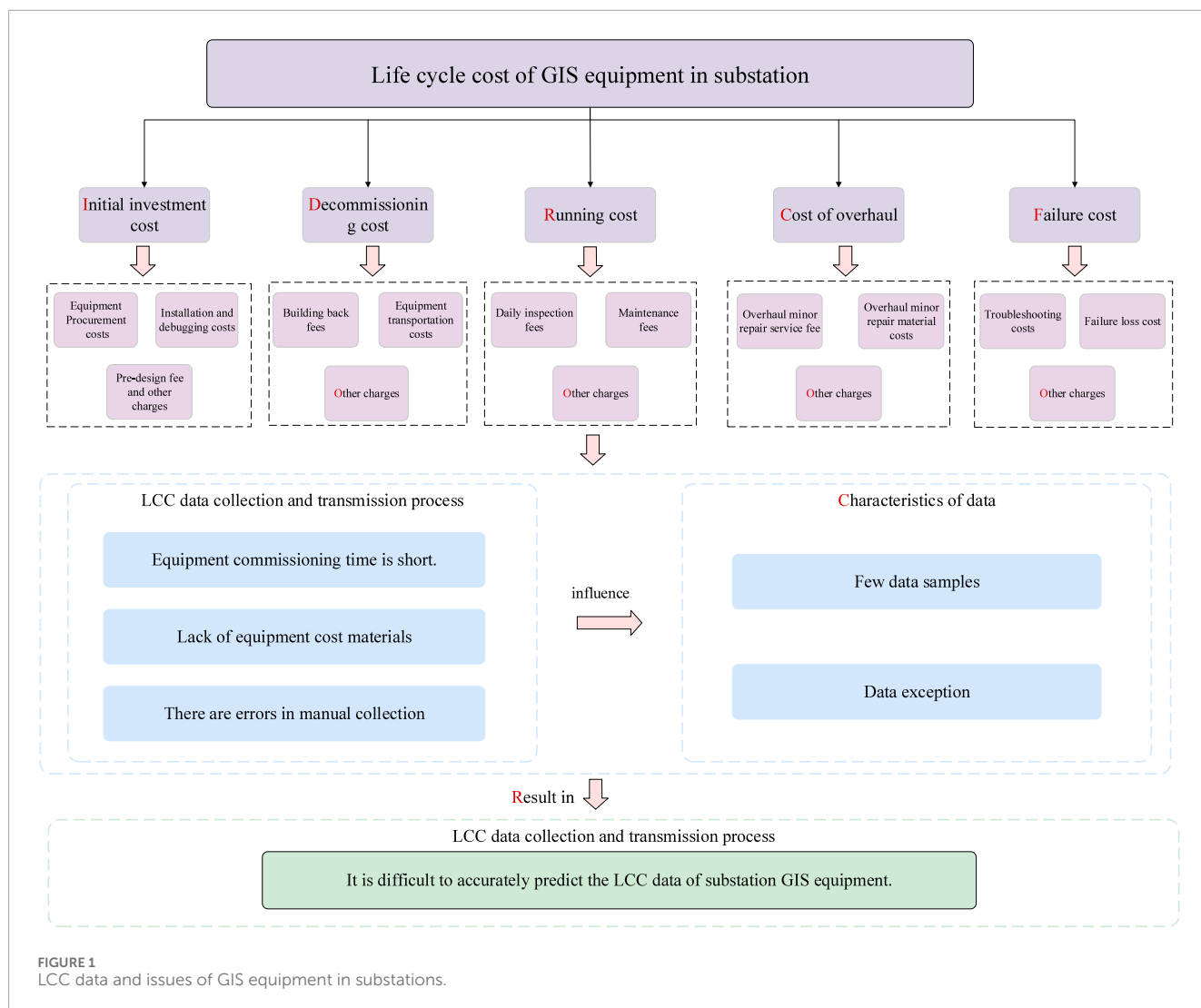
From the perspective of actual data collection on variable annual costs, GIS equipment in substations only began to be widely used in China in the 21st century, and related data collection activities started relatively late. As a result, the quantity of LCC data collected across different years is limited, leading to a sparse overall LCC data sample. Additionally, in the early stages of LCC data collection and transmission, manual recording methods were commonly relied upon. Consequently, issues such as insufficient cost materials, data loss, and recording errors have occurred, resulting in anomalies within the collected LCC sample data.

In summary, these data issues have created significant challenges for accurately calculating LCC data. Therefore, it is necessary to identify an appropriate data prediction method based on the characteristics of the LCC data for GIS equipment in substations. This method should aim to expand the data sample size while minimizing the impact of anomalous data on the predictions.

2.2 Limitations of existing data prediction methods and proposed improvements

As discussed in Section 1.1, to achieve accurate predictions of LCC data for GIS equipment in substations, it is essential first to expand the sample size as much as possible to address the issue of limited data samples. Then, the characteristics of the data samples should be analyzed to resolve the issue of data anomalies.

Currently, under the condition that sample data exhibit certain regular variations, an effective approach to address the issue of limited sample data for prediction is to introduce similar data



of the same scale as the existing sample data, thereby expanding the original dataset (Zhang et al., 2022; Wang et al., 2022; Yang et al., 2023; Cheng et al., 2022). However, considering that the data characteristics of the existing samples and the similar data may not be entirely consistent, an attention mechanism with focused characteristics is employed to quantify the influence weights between the existing sample data and similar data. The weighted data is then used as the expanded sample dataset (Tao et al., 2018). However, the attention mechanism requires manually assigned data feature vectors to process and analyze the data around these vectors. Given that LCC data for GIS equipment in substations belongs to cost data, which is significantly influenced by the actual operational conditions of the equipment, the data distribution pattern is complex, and the data features are not prominent. Therefore, manual feature extraction lacks a solid basis and accuracy. As a result, directly using the attention mechanism to obtain influence weights between data may lead to low accuracy. If a method could calculate vector weights between data to automatically extract feature vectors, enhancing the reliability of feature extraction results, it would avoid the subjectivity associated with manual feature vector extraction. This would provide an objective basis

for quantifying the influence weights between existing sample data and similar data, thus improving the accuracy of data predictions.

Regarding the issue of anomalous data, after performing data preprocessing based on data distribution characteristics, it is important to consider that GIS equipment LCC data is collected and arranged in chronological order, making it sequential data. Therefore, a Recurrent Neural Network (RNN), which can capture temporal relationships, is employed for time-series data prediction (Ye et al., 2023; Li and Lin, 2018). In processing time-series data such as LCC data, RNN adjusts neural network parameters by calculating gradients according to the backpropagation algorithm and the chain rule (Kong et al., 2020) to minimize the loss function (Kong et al., 2020), thereby reducing the overall error in the prediction results. However, according to the chain rule, each parameter's gradient expression in RNN is a product of terms, which often causes the calculated values to converge exponentially toward zero or infinity as the period increases. This leads to the gradient vanishing or exploding problem in the LCC data, raising concerns about the accuracy of the predictions. The fundamental cause of gradient vanishing or exploding in RNN

lies in the linear nature of its information transfer structure. Therefore, by optimizing the internal structure of the network to selectively retain long-span information, the non-linear relationship between data points can be enhanced, suppressing the occurrence of gradient vanishing or exploding during the network training and updating process.

In summary, to achieve accurate LCC data predictions for GIS equipment in substations, we can first introduce a self-attention mechanism to enhance the feature representation of existing sample data, quantifying the influence weights between existing sample data and similar data, thereby expanding the sample size. Then, a GRU model can be employed, leveraging its gated structure to suppress the gradient vanishing or exploding issues, further improving the accuracy of the prediction results and enabling precise LCC data predictions.

3 SA-GRU model for LCC data prediction

This chapter details the development of the SA-GRU model for LCC prediction, targeting two core challenges in substation GIS equipment LCC forecasting: the insufficiency of sample sizes and the presence of gradient anomalies in time-series data. The former challenge is addressed by quantifying the influence weights of data from similar substations through the self-attention mechanism, thereby expanding the sample dataset. The latter challenge is alleviated via the gated structure of the GRU, which suppresses gradient vanishing and exploding phenomena during model training.

3.1 Sample expansion based on self-attention mechanism

Given the limited number of existing sample data, it is feasible to introduce LCC-similar data from GIS equipment in substations of the same voltage level and similar environments to expand the sample size. However, since similar data cannot fully represent the characteristics of the existing sample data, directly adding similar data to the existing samples for cost prediction would fail to achieve precise LCC calculations. Therefore, it is necessary to apply weighting to similar data based on the characteristics of the existing samples. This approach addresses the issue of reduced prediction accuracy caused by the incomplete consistency between similar data and existing sample data characteristics.

Due to the complex distribution patterns of LCC data for GIS equipment in substations, identifying correlations between data for weighting purposes requires transforming both existing sample data and similar data into matrix form. Different years are used as the row labels of the data, and various types are used as the column labels of the matrix. Based on this, a matrix representation of the existing sample data is constructed as follows:

$$H_{LCC} = \begin{bmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & & \vdots \\ h_{t1} & \cdots & h_{tn} \end{bmatrix}. \tag{1}$$

In this expression, H_{LCC} represents the constructed matrix of existing sample data; h_{tn} denotes the LCC data of the n -th type in the t -th year for GIS equipment; $n = 1, 2, 3 \dots, N$; $t = 1, 2, 3 \dots, T$.

Considering that the similar data includes LCC data from multiple GIS equipment in substations of the same scale, the matrix is divided into sub-blocks by different substations as column references. Each sub-matrix is constructed in a manner consistent with the matrix representation of the existing sample data. To facilitate correlation comparisons between different data sets, the existing sample data is included as a reference value within the similar data, so that the matrix form of the similar data can be represented as follows:

$$L_{LCC} = [H_{LCC}, H_{LCC}^1, H_{LCC}^2, \dots, H_{LCC}^{m-1}, H_{LCC}^m]. \tag{2}$$

In this expression; L_{LCC} denotes the similarity data sub-matrix composed of the LCC data from the m -th GIS equipment in substations of the same scale. The number of rows and columns in Hm LCC matches those of the existing sample data matrix.

Based on the matrix constructions in Equations 1 and 2, it can be seen that the task of identifying correlations between existing sample data and similar data has been transformed into a problem of determining correlations between matrix column vectors (Zhang, 2021). According to the attention mechanism's solving algorithm (Yin et al., 2021), the correlation between matrix column vectors can be obtained using an attention scoring function, specifically:

$$\alpha_m = \text{Softmax}(s(H_{LCC}^m, H_{LCC})). \tag{3}$$

In this expression, α_m represents the weight coefficient matrix between the m -th similar data sub-matrix and the existing sample data matrix, indicating the correlation between matrices; Softmax is a normalization function; and $s(\cdot)$ is the attention scoring function. Considering the long time sequence of LCC data and the suitability of the dot-product model for calculating vector weight coefficients between matrices, the attention scoring function in Equation 3 employs a scaled dot-product model.

In solving matrix correlations using the scoring function, the attention scoring function performs calculations directly based on the matrix elements. Given that the LCC data of GIS equipment has inconspicuous data features, the scoring results of the attention function may lack accuracy. Therefore, an algorithm based on the self-attention mechanism (Yin et al., 2021) is introduced to determine the weight relationships among elements within the H_{LCC} matrix, thereby enhancing data features. When solving for weight relationships with the self-attention mechanism, to better capture the correlations between any two elements in the matrix, the H_{LCC} matrix needs to be transformed into three matrices through linear mapping: the LCC data query vector matrix Q_{LCC} , the LCC data key vector matrix K_{LCC} , and the LCC data value vector matrix V_{LCC} . Here, matrices Q_{LCC} and K_{LCC} share the same dimensions and are used to calculate the weight coefficients between column vectors of matrix H_{LCC} , while matrix V_{LCC} is used to compute the output by combining them with the weight coefficients. At this point, the weight coefficients between Q_{LCC} and K_{LCC} represent the weight coefficients between elements of the H_{LCC} matrix, and these coefficients can also be obtained using a scoring function:

$$A_{LCC} = \text{Softmax}(s(K_{LCC}, Q_{LCC})). \tag{4}$$

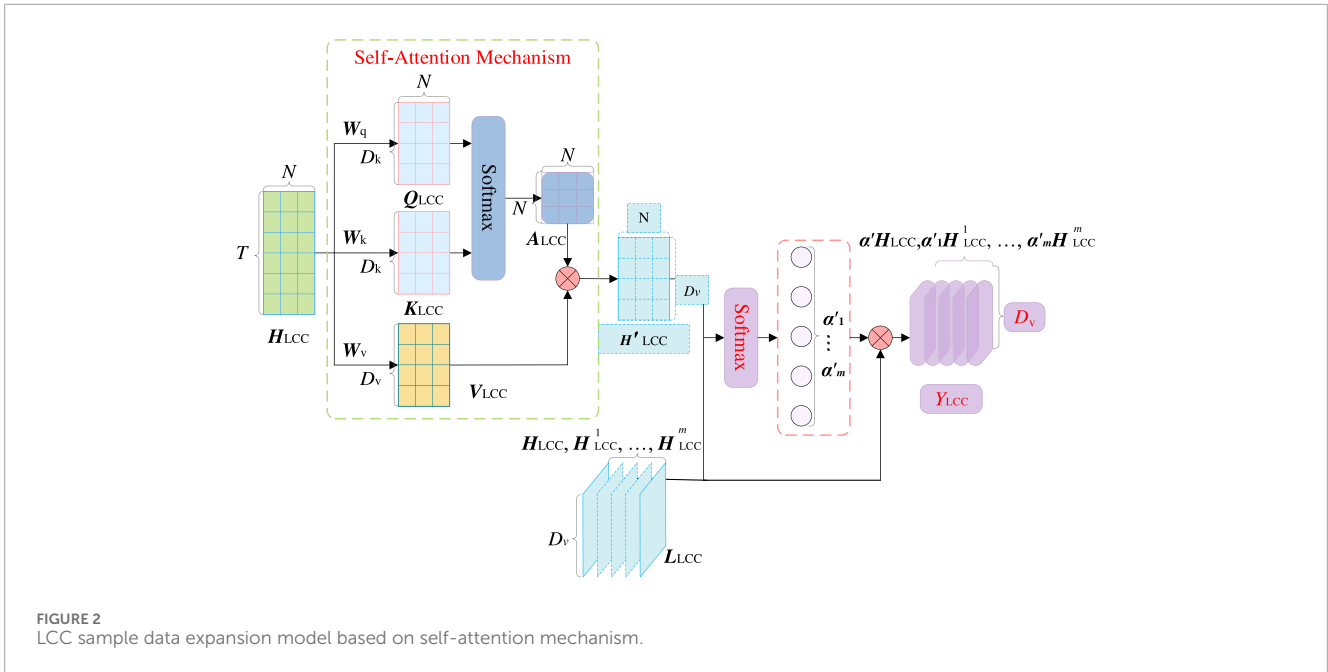


FIGURE 2 LCC sample data expansion model based on self-attention mechanism.

In this expression, A_{LCC} represents the weight coefficient matrix of the existing sample data; Softmax is a normalization function, and $s(\cdot)$ is the attention scoring function. Since both matrices Q_{LCC} and K_{LCC} are also sequential data and involve calculating vector weight coefficients between matrices, the scaled dot-product formula is similarly used in Equation 4.

Since the weight coefficient matrix A_{LCC} obtained from Equation 4 only represents the correlations between matrix elements, to obtain a matrix with enhanced data features, it is necessary to perform matrix multiplication between A_{LCC} and V_{LCC} . This results in the data feature matrix:

Since the weight coefficient matrix A_{LCC} obtained from Equation 4 only represents the correlations between matrix elements, to obtain a matrix with enhanced data features, it is necessary to perform matrix multiplication between A_{LCC} and V_{LCC} , resulting in the data feature matrix:

$$H'_{LCC} = V_{LCC}A_{LCC} \tag{5}$$

In this expression, H'_{LCC} is the data feature matrix, with the same rows and columns as matrix H_{LCC} .

Furthermore, by substituting the enhanced data feature matrix H'_{LCC} of Equation 5 into Equation 3 for calculation, the weight coefficient matrix α'_m between the m -th similar data sub-matrix and the data feature matrix can be obtained. On this basis, to reduce the impact of the partial inconsistency between the data features of matrices H'_{LCC} and H^m_{LCC} on prediction accuracy, matrix α'_m is multiplied by matrix H^m_{LCC} . This yields the LCC data matrix for GIS equipment in substations after expanding the sample data in Equation 6.

$$Y_{LCC} = [\alpha'_1 H^1_{LCC}, \dots, \alpha'_m H^m_{LCC}]. \tag{6}$$

In this expression, Y_{LCC} represents the expanded sample data matrix.

Thus, the final LCC sample data expansion model based on the self-attention mechanism is constructed, as shown in Figure 2.

3.2 Data prediction based on GRU

Since there are anomalies in the LCC data, directly using the expanded sample data for prediction would still affect the accuracy of data predictions. Therefore, to fundamentally improve prediction accuracy, it is necessary to handle the anomalies in the LCC data. Given that anomalies in LCC data only appear at specific points in the time series, classifying them as point anomalies (Li et al., 2023), a weighted moving average method is employed to process the expanded sample data to mitigate the impact of point anomalies in the original time series:

$$\begin{cases} Y_{LCC}^{t+1} = \frac{\sum_{i=1}^t w_i Y_{LCC}^i}{\sum_{i=1}^t w_i} \\ Y'_{LCC} = [Y_{LCC}^1, \dots, Y_{LCC}^T]^T \end{cases} \tag{7}$$

In this expression, Y_{LCC}^i represents the expanded sample data value for the i -th year; w_i denotes the data weight for the i -th year; $i \in [1, T]$; t is the year for which the weighted moving average method is used for data processing; and Y'_{LCC} is the expanded sample data matrix after processing.

On this basis, considering that the recurrent structure of the RNN algorithm can effectively capture the overall trend of time-series data, the RNN algorithm is introduced to predict the processed expanded sample data (Yang et al., 2023). Matrix Y'_{LCC} is used as the data input, and the RNN algorithm performs network training and prediction on the input data through its gradient calculation process.

$$\begin{aligned} \frac{\partial L_t}{\partial V} &= \sum_{k=0}^t \frac{\partial L_k}{\partial y_k} \frac{\partial y_k}{\partial h_k} \left(\prod_{j=k-1}^t \frac{\partial h_j}{\partial h_{j-1}} \right) \frac{\partial h_k}{\partial V} \\ &= \sum_{k=0}^t \frac{\partial L_k}{\partial y_k} \frac{\partial y_k}{\partial h_k} \left(\prod_{j=k-1}^t \tan h' \cdot U \right) \frac{\partial h_k}{\partial V}. \end{aligned} \tag{8}$$

In this expression, L_t represents the loss function of the RNN network at time t ; y_k is the output of the RNN network at time k ; h_j is the hidden state of the RNN network's hidden layer at time j ; \tanh is the activation function of the RNN hidden layer; V is the weight parameter from the RNN hidden layer to other layers; and U is the weight parameter between hidden layers of the RNN.

From the linear multiplicative relationship in Equation 8, it can be seen that if $\tanh \cdot U$ is less than 1, the gradient will converge exponentially to zero as the time span of the input LCC data increases. Conversely, if $\tanh \cdot U$ is greater than 1, the gradient will tend towards infinity as the time span increases. Therefore, during gradient calculations over long time spans, gradient vanishing or exploding issues may arise, resulting in decreased prediction accuracy. To address this issue, considering that the LCC data for GIS equipment in substations does not belong to a large dataset, and to facilitate computation, a GRU gating structure is added to the RNN algorithm's architecture. This modification mitigates the gradient vanishing and exploding problems caused by the linear multiplicative relationship.

The network structure with the added GRU gating structure can be expressed as follows in Equation 9:

$$\begin{cases} \mathbf{r}_t = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r) \\ \mathbf{z}_t = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z) \\ \mathbf{h}'_t = \tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h (\mathbf{h}_{t-1} \odot \mathbf{r}_t) + \mathbf{b}_h) \\ \mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \mathbf{h}'_t \end{cases} \quad (9)$$

In this expression, \mathbf{r}_t is the reset gate at the current time step; \mathbf{z}_t is the update gate at the current time step; \mathbf{h}'_t represents the candidate state information at the current time step; \mathbf{x}_t is the input information at the current time step; \mathbf{h}_t is the hidden state information passed to the next time step; \mathbf{h}_{t-1} is the hidden state information from the previous time step; σ is the sigmoid activation function; \tanh is the activation function; \odot denotes the Hadamard product, which is the element-wise multiplication of matrices; \mathbf{W}_r , \mathbf{W}_z , \mathbf{W}_h , \mathbf{U}_r , \mathbf{U}_z , \mathbf{U}_h are the weight parameter matrices corresponding to the reset gate, update gate, and candidate state; \mathbf{b}_r , \mathbf{b}_z , \mathbf{b}_h are the bias parameter matrices for the reset gate, update gate, and candidate state, respectively.

Thus, the multiplicative relationship expression for gradient calculation in the GRU structure becomes:

$$\begin{cases} \prod_{j=k-1}^t \frac{\partial h_j}{\partial h_{j-1}} = \prod_{j=k-1}^t \left((1 - z_j) + (U_z \delta_{z_j} + U_h (\delta_{h_j} r_j)) + (U_r \delta_{r_j}) \right) \\ \delta_{z_j} = \frac{\partial L_j}{\partial z_j} = \frac{\partial L_j}{\partial h_j} [h_j - h_{j-1}] z_j (1 - z_j) \\ \delta_{r_j} = \frac{\partial L_j}{\partial r_j} = \delta_{h_j} (U_h h_{j-1}) r_j (1 - r_j) \\ \delta_{h_j} = \frac{\partial L_j}{\partial h_j} = \frac{\partial L_j}{\partial h_j} z_j (1 - h_j^2) \end{cases} \quad (10)$$

In this expression: δ_{z_j} is the partial derivative of the loss function concerning the update gate; δ_{r_j} is the partial derivative of the loss function concerning the reset gate; δ_{h_j} is the partial derivative of the loss function concerning the hidden state information.

By substituting the multiplicative terms in Equation 10 into Equation 8, it can be observed that the original linear multiplicative

relationship in Equation 8 has been replaced by a nonlinear relationship. Consequently, the gradient vanishing and exploding issues in data prediction are resolved. At this point, the processed expanded sample data matrix can be used as input to the network, and data prediction is performed through the GRU network.

After obtaining the prediction values from the GRU network, the most commonly used metrics in time series prediction—Root Mean Square Error (RMSE) in Equation 11, Mean Absolute Error (MAE) in Equation 12, and R^2 (R-Square) in Equation 13—are selected for error validation:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (11)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%, \quad (12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (13)$$

In these expressions: y_i represents the true value of the sample; \hat{y}_i denotes the predicted value; n is the number of true values in the sample, \bar{y} represents the mean of the predicted values.

3.3 Data prediction process

Combining the above analyses of the self-attention mechanism and GRU, the process begins by applying weight processing to similar data based on the self-attention mechanism to obtain expanded sample data. Then, the expanded sample data is used to train model parameters with the GRU model. Finally, the LCC data prediction results are output, and the accuracy of the prediction results is validated. The complete LCC data prediction workflow in Figure 3 is as follows.

4 Example verification and analysis

4.1 Actual conditions of the substation

Taking the 110 kV GIS equipment of a substation in Henan Province as an example, this substation is selected as the sample substation, and the SA-GRU model proposed in this paper is applied to predict the LCC data of GIS equipment in this sample substation. The substation currently has two main transformers, and the 110 kV side distribution device uses outdoor GIS equipment. It is located in an E-grade pollution area. The basic information on the voltage level, environment, and scale of the sample substation is shown in Table 1.

To demonstrate that the SA-GRU model proposed in this paper is suitable for predicting the LCC data of GIS equipment in substations, three similar substations are introduced to expand the LCC data samples based on the basic conditions, such as the voltage level, electrical scale, and surrounding environment of the example substation. The basic information of the introduced similar substations is shown in Table 2.

Comparing the basic information of substations in Tables 1 and 2 reveals that the three introduced substations have the same voltage level as the sample substation, with altitudes below 1,000 m, are

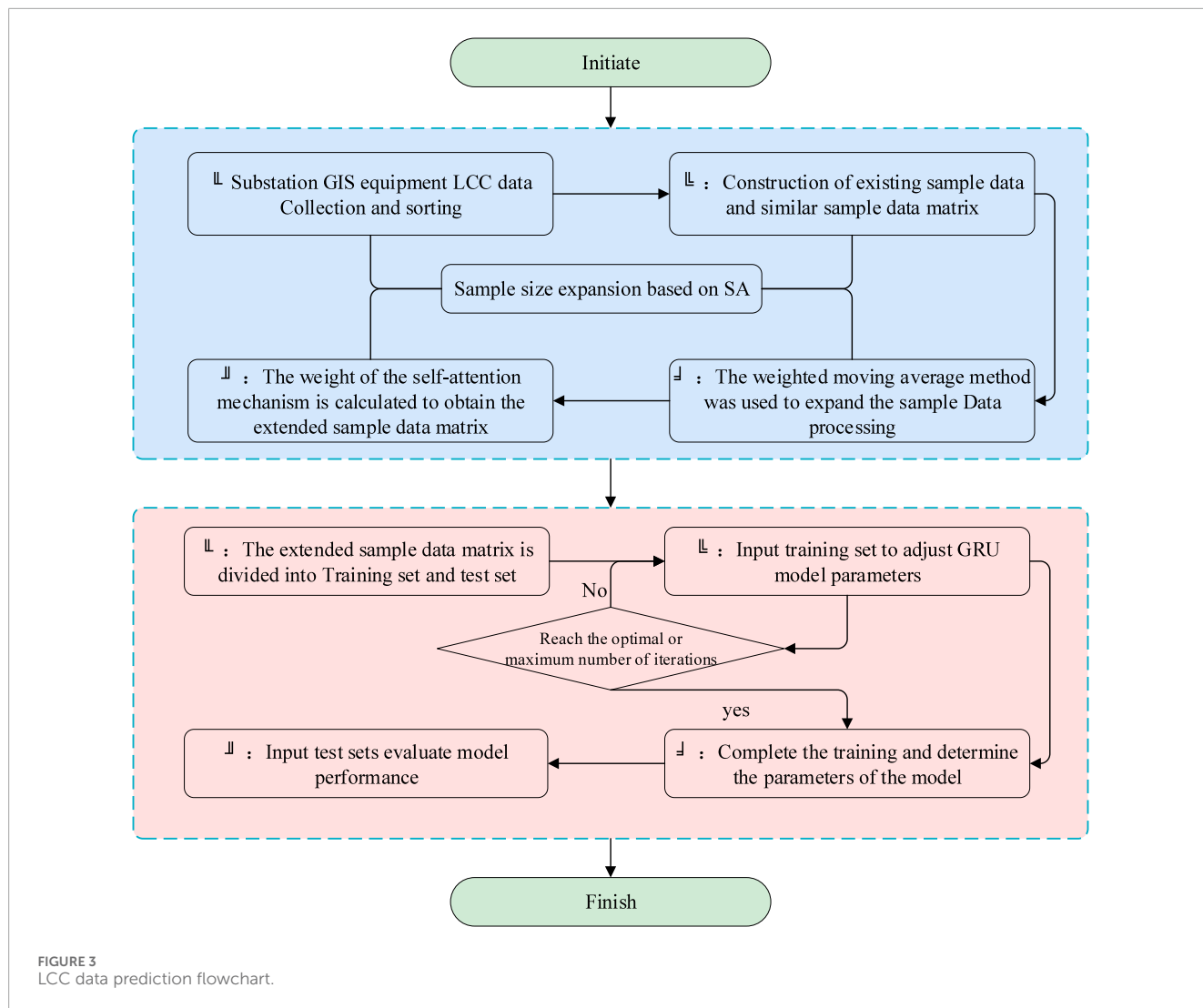


TABLE 1 Basic information of the sample substation.

Item	Numerical value	Unit
Voltage level	220	kV
Number of main converters	2	Set number
Principal variable capacity	180	MVA
110 kV incoming and outgoing lines	12	Circuit number
GIS equipment layout	Outdoors	—
Altitude	<1,000	m
Pollution grade	E	Level

located in an E-grade pollution area, and all GIS equipment is arranged outdoors. Substations 1 and 2 have one more main transformer than the sample substation, while Substation 3 has the

same number of main transformers as the sample substation. In terms of incoming and outgoing line circuits, Substation 1 has the same number as the sample substation, Substation 2 has 4 more, and Substation 3 has 2 fewer circuits. Among these, Substation 3 has the highest similarity to the sample substation in terms of basic information, while Substations 2 and 1 have lower similarity.

After identifying the sample substation and similar substations, raw LCC data were obtained through the PMS and ERP system platforms and multidimensional lean reports. This data collection primarily focused on cost-related data, including voltage level, service life, asset original value, routine maintenance and repair material costs, major repair material costs, maintenance fees, labor costs, and other expenses. As of 31 December 2023, a total of 886 LCC sample data entries for GIS equipment were collected for the sample substation. Additionally, 4,101 LCC sample data entries were gathered for GIS equipment from the three similar substations, facilitating the expansion of LCC data samples.

A portion of the collected sample data on cost-related information for 110 kV GIS equipment is shown in [Table 3](#).

To clearly present the data foundation of this study, [Table 4](#) lists representative samples of the LCC time-series data collected

TABLE 2 Basic information of similar substations.

Item	Substation 1	Substation 2	Substation 3
Voltage level	220 kV	220 kV	220 kV
Number of main converters/unit	3	3	2
Main variable capacity/MVA	240	180	180
110 kV incoming/outgoing line/return	12	16	10
GIS equipment layout	Outdoors	Outdoors	Outdoors
Altitude m	<1,000	<1,000	<1,000

TABLE 3 Sample of cost-related data for 110 kV GIS equipment (partial).

Serial number	Operation age limit	Original value of assets	Maintenance and repair materials cost	Cost of major repair materials	Repair charge	Labor cost
1	38	143435.27	414.30	1	2959.91	2631.77
2	32	359302.77	2100.87	10676.74	588.70	35819.53
3	18	1557504.77	1515.03	1515.03	3680.27	15061.98
4	23	363904.65	222.28	0	4864.15	5059.33
5	4	171930.08	708.70	0	0	4484.22
6	12	873264.77	134.56	0	1139.65	2584.88
7	2	124368.33	7468.46	5193.79	7509.21	1856.92
8	9	645125.57	8415.71	9456.00	1180.20	53565.84

from a 110 kV GIS unit at a substation in Henan Province. The dataset contains cost records spanning the equipment’s entire life cycle, reflecting the continuity and complexity characteristic of time-series data.

The GRU model was configured with the following hyperparameters: 64 hidden units, a look-back window (time steps) of 10, a dropout rate of 0.3, a batch size of 16, the Adam optimizer with an initial learning rate of 0.001, and 200 training epochs. These hyperparameters were selected through preliminary tuning based on validation performance, and early stopping was employed during training to prevent overfitting.

4.2 Weight calculation results based on self-attention mechanism

Based on the LCC sample data expansion model using the self-attention mechanism, the influence weights of LCC data for GIS equipment in different substations can be obtained. A comparison of these weights with those obtained using the direct attention mechanism is shown in Figure 4.

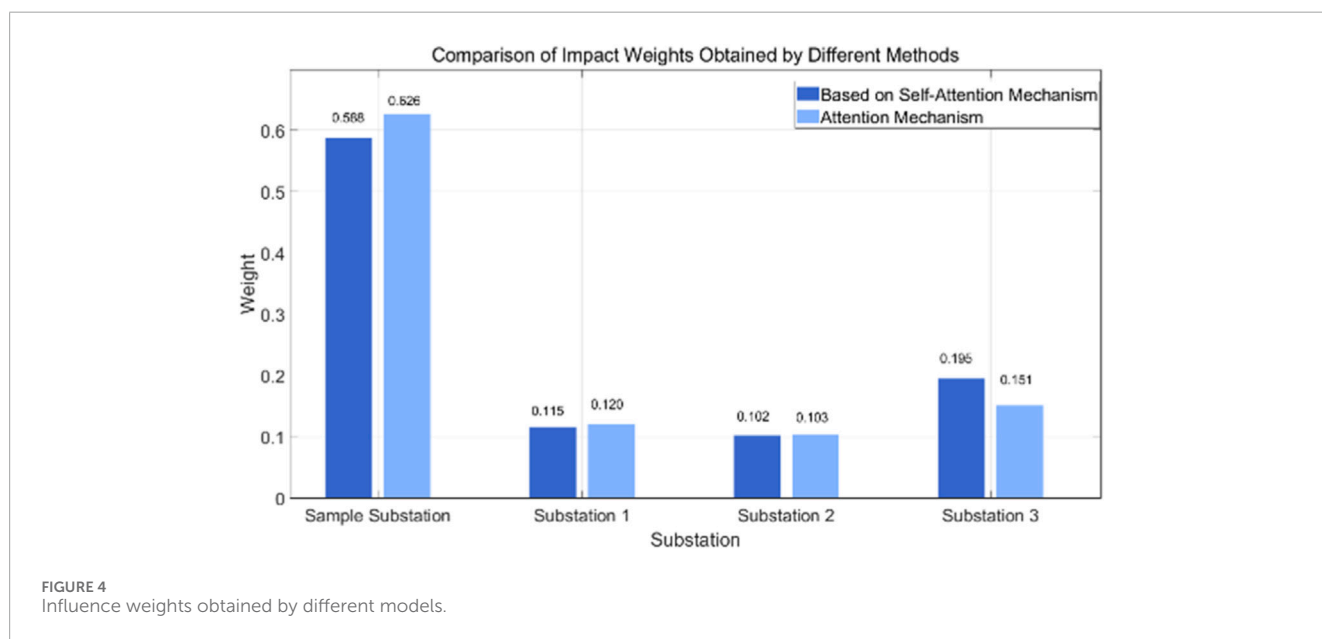
As shown in Figure 4, the influence weights based on the self-attention mechanism for the sample substation, Substation 1,

Substation 2, and Substation 3 are 0.588, 0.115, 0.102, and 0.195, respectively. The weight results indicate that, apart from the sample substation, Substation 3 has the highest weight, likely due to the similarity of its main transformer count, main transformer capacity, and the number of incoming and outgoing circuits with those of the sample substation, differing only by 2 fewer circuits. The weights of Substation 1 and Substation 2 are nearly 50% lower than that of Substation 3, likely because Substation 1 has a higher main transformer count and capacity, and Substation 2 has a higher main transformer count and a larger number of incoming and outgoing circuits than the sample substation. These quantitative differences lead to higher LCC data values for Substations 1 and 2. Substation 2 has the lowest weight, possibly because the number of incoming and outgoing circuits on the 110 kV side has a greater impact on the LCC data values.

Comparing the influence weights of the attention mechanism in Figure 5, it can be seen that the weight of the sample substation based on the self-attention mechanism is reduced by 0.038 compared to the attention mechanism. This may be because the attention mechanism directly uses the raw sample data as sample features, whereas the self-attention mechanism extracts sample features from the raw data, leading to some differences between the two types of sample features. Except for Substation 3, the influence

TABLE 4 LCC time-series data collected from a 110 kV GIS unit at a substation in Henan Province.

Serial number	LCC cost (Yuan)	Serial number	LCC cost (Yuan)
1	10253.92	11	21117.94
2	18594.72	12	22019.09
3	17078.66	13	18771.44
4	20873.63	14	24503.27
5	21922.14	15	25276.93
6	25626.14	16	28541.20
7	18292.75	17	28401.98
8	23898.47	18	30419.14
9	24465.10	19	30956.61
10	27341.49	20	25864.47



weights of the other similar substations based on the self-attention mechanism are all reduced. This may be because the LCC data features of Substation 3 are closer to the features extracted by the self-attention mechanism. Additionally, Substation 2 still has the smallest weight result, while Substation 3 still has the largest.

4.3 Comparative analysis of training results from different models and methods

Based on the typical lifespan of GIS equipment in substations, 40 years is selected as the overall time frame for LCC, with predictions made annually. Considering that the LCC data for substation GIS equipment, even after sample expansion, still

constitutes a small dataset, no validation set is created. Instead, the sample data processed in Equation 7 is split into a training set and a test set in a traditional 7:3 ratio. The training set is used to train model parameters, and the test set is used to validate the model's training results. To maintain continuity in the time-series data during training, a segmented approach is adopted for dividing the dataset: the first 32 consecutive years of data are used for model parameter training, and the last 8 consecutive years of data are used for validation. The divided training set is fed as input to the GRU model, with only LCC values as the output. Considering the components of the model in this paper, the SA-GRU model, ATT-GRU model (Attention combined with GRU model), and GRU model are selected for a comparative analysis of the training results from different models. The training results of the different

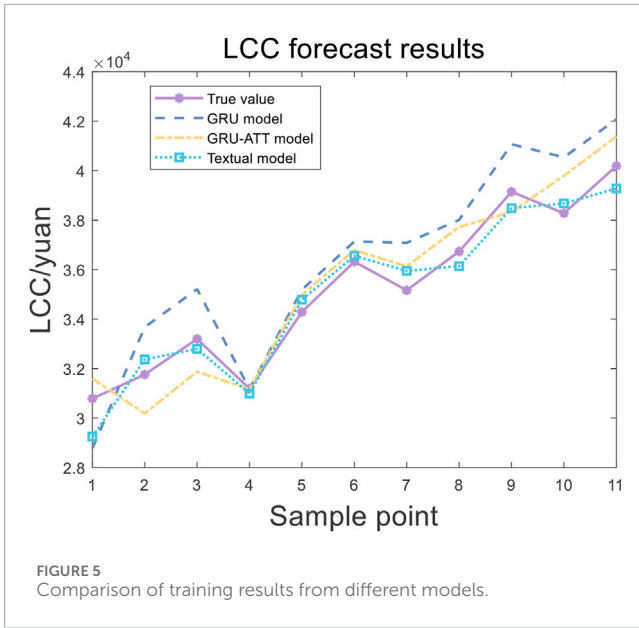


TABLE 5 RMSE, MAE, and R^2 metrics results for different models and different methods.

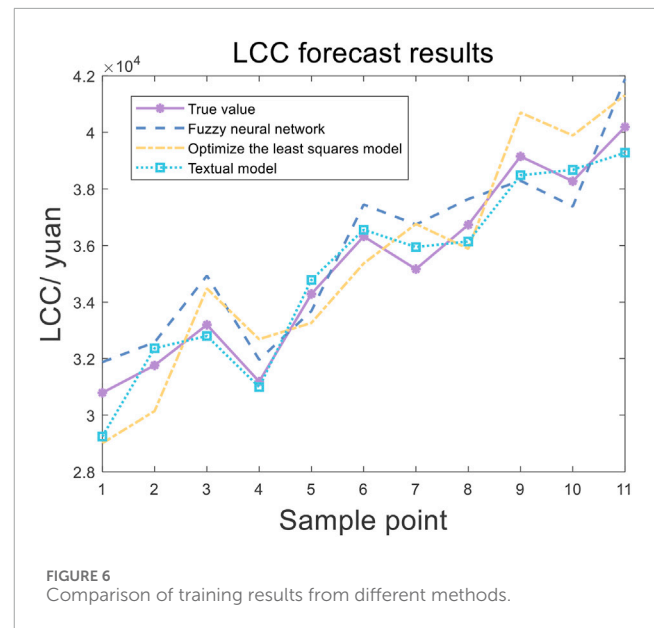
Prediction model	RMSE/元	MAPE	R^2
GRU model	1677.33	3.83%	83.32%
The ATT-GRU model	1039.18	2.35%	91.78%
Optimize partial least squares	1388.11	3.37%	87.31%
The fuzzy neural network method	1160.57	2.73%	85.68%
Textual method	715.93	1.54%	94.86%

models on the LCC data of substation GIS equipment are shown in Figure 5, and the evaluation metrics RMSE, MAE, and R^2 are presented in Table 4.

As seen in Figure 5, from an overall training perspective, while the GRU model's general trend aligns with the true values, the discrepancies between its predictions and the actual values are noticeably larger than those of the SA-GRU and ATT-GRU models. The ATT-GRU model, which is an improved version of the GRU model with the inclusion of similar data during modeling, shows certain advantages in small-sample data prediction. Although its performance is not as accurate as the SA-GRU model, the overall prediction results are still better than the GRU model, closely approximating the true values, with only slight deviations at certain sample points and towards the end of the training. The SA-GRU model used in this paper provides results that are even closer to the true values compared to the ATT-GRU model. This improvement is due to the SA-GRU model's use of a self-attention mechanism to enhance sample data features, which increases the weights assigned to similar substations that closely match the sample data features. This adjustment reduces the negative impact on prediction accuracy caused by inconsistencies between the features of similar data and sample data.

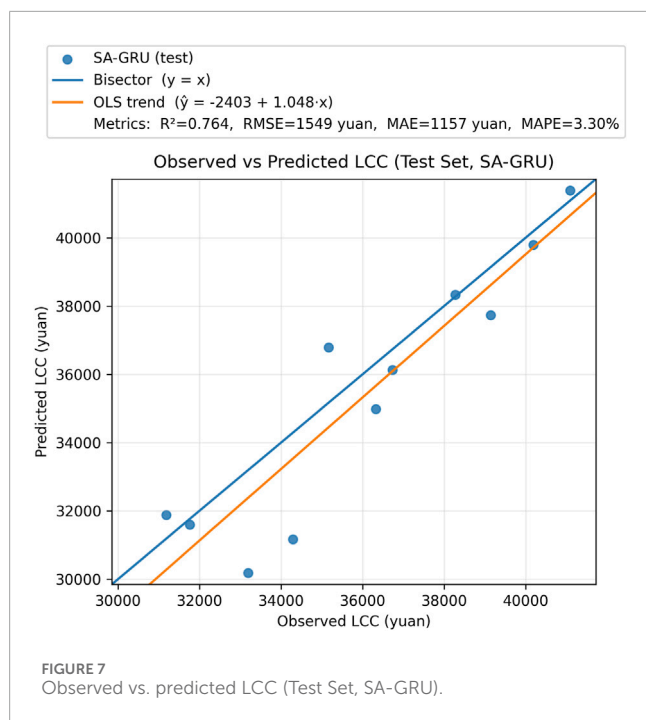
From Table 5, it can be observed that the RMSE of the SA-GRU model is 715.93 yuan, the MAPE is 1.54%, and the R^2 value is 94.86%; the RMSE of the ATT-GRU model is 1039.18 yuan, the MAPE is 2.35%, and the R^2 value is 91.78%; while the GRU model's RMSE is 1677.33 yuan, the MAPE is 3.83%, and the R^2 value is 83.32%. Compared to the ATT-GRU model, the SA-GRU model improves the RMSE accuracy by 31.11% and the MAPE accuracy by 34.47%, with the R^2 value increasing from 91.78% to 94.86%.

To validate the accuracy of the proposed method for LCC data prediction, the proposed method was compared with the optimized partial least squares method and the fuzzy neural network method mentioned in the introduction. The training results are shown in Figure 6.



According to the evaluation index calculation formulas, the RMSE, MAE, and R^2 results of different prediction methods are presented in Table 4. Combined with Figure 6, it is evident that among time-series prediction methods, the SA-GRU model demonstrates superior training performance and evaluation results compared to other methods. The optimized partial least squares (PLS) method is significantly influenced by numerical fluctuations during training, resulting in an inability to accurately capture data trends. Although the fuzzy neural network method mitigates the impact of numerical fluctuations on training, it shows weaker capability in capturing data features. These limitations cause both methods to yield less accurate prediction results than the proposed method in this paper. As shown in Table 4, compared to the optimized PLS method, the proposed method reduces the RMSE by 672.18 yuan, decreases the MAPE by 1.83%, and improves the R^2 by 7.55%. Compared to the fuzzy neural network method, the proposed method reduces the RMSE by 444.64 yuan, decreases the MAPE by 1.19%, and improves the R^2 by 9.18%.

To visualize the predictive performance of the best model, Figure 7 plots observed versus predicted LCC on the held-out test set. The 45° bisector ($y = x$) is shown as the ideal



reference, together with an ordinary-least-squares (OLS) fit of predicted on observed values. The legend reports R^2 , RMSE, MAE, and MAPE to complement the visual assessment.

In Figure 7, points cluster around the bisector. The OLS fit $\hat{y} = -2403 + 1.048x$ has a slope close to 1 and a small intercept, indicating limited systematic bias. Quantitatively, on the test set we obtain $R^2 = 0.764$, RMSE = 1,549 yuan, MAE = 1,157 yuan, and MAPE = 3.30%, which is consistent with Table Y. The slope slightly above 1 suggests a mild overestimation at higher LCC levels (and slight underestimation for low-cost years), but the deviation is small relative to the overall range. Overall, SA-GRU provides accurate and decision-grade LCC forecasts.

In summary, the SA-GRU model proposed in this paper can effectively improve the accuracy of LCC data prediction for GIS equipment in substations.

5 Conclusion

Considering the limited sample size and presence of outlier data in the GIS equipment LCC data of substations, this study proposes an SA-GRU-based prediction method for LCC data. Taking the 110 kV GIS equipment of a substation in Henan Province as an example, the conclusions are as follows:

1. The SA-GRU model outperforms the Attention-GRU model and the GRU model in terms of RMSE, MAPE, and R^2 evaluation metrics. Models incorporating self-attention mechanisms also demonstrate superior performance compared to those using only attention mechanisms, confirming the necessity of using self-attention to process data from similar substations.
2. Compared with traditional optimized partial least squares (PLS) and fuzzy neural network methods, the proposed

method reduces the RMSE by 672.18 yuan and 444.64 yuan, decreases the MAPE by 1.83% and 1.19%, and increases the R^2 by 7.55% and 9.18%, respectively. These results validate the effectiveness of the proposed method in improving the accuracy of GIS equipment LCC data prediction and addressing the challenges in accurately forecasting LCC data for substations.

3. The proposed method quantifies the influence weights of data from the sample substation and similar substations to obtain weighted data from similar substations, thereby expanding the sample size. However, the current approach to weight calculation does not consider the qualitative characteristics of individual substations. Future research could focus on quantifying the scale and environmental characteristics of substations and developing appropriate methods to further enhance the reliability of weight calculations for substation interactions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

WX: Writing – original draft, Writing – review and editing. CA: Conceptualization, Data curation, Resources, Writing – review and editing. TS: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing. PD: Data curation, Formal Analysis, Investigation, Software, Writing – review and editing. YZ: Writing – review and editing, Methodology, Resources, Visualization. JZ: Formal Analysis, Project administration, Supervision, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the State Grid Technology Project (B717L028K136).

Acknowledgements

This paper thanks all the authors for their dedicated work and the financial support of various institutions. And thanks for the comments from all reviewers.

Conflict of interest

Authors WX, CA, TS, PD, YZ, and JZ were employed by State Grid Henan Electric Power Company.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

References

- Cai, Y., Liu, L., and Cheng, H. (2011). A review of life-cycle cost (LCC) technology applications in power systems. *Power Syst. Prot. Control* 39 (17), 149–154. doi:10.7667/fj.issn.1674-3415.2011.17.029
- Cheng, K., Peng, X., and Xu, Q. (2022). Short-term wind-farm forecasting feature selection and multi-level deep transfer learning. *High. Volt. Eng.* 48 (2), 497–503. doi:10.13336/j.1003-6520.hve.20210032
- Du, J. (2013). "Life-cycle cost analysis and management of underground substations based on fuzzy theory," in *Master's thesis*. Shanghai Jiao Tong University.
- Kong, X., Luo, J., and Zhang, Q. (2020). Quality-related fault detection using orthogonal signal correction and efficient partial least squares. *Control Decis.* 35 (5), 1167–1174. doi:10.13195/j.kzyjc.2018.0708
- Li, J., and Lin, Y. (2018). Time-series forecasting with multi-time-scale RNN. *Comput. Appl. Softw.* 35 (7), 33–37. doi:10.3969/j.issn.1000-386x.2018.07.006
- Li, X., Tian, F., and Guo, Y. (2023). Key technologies of health management and intelligent O&M for power equipment in the new power system. *Power Syst. Technol.* 1–18. doi:10.13335/j.1000-3673.pst.2022.2451
- Liu, L., Jiang, G., and Sun, Y. (2020a). LCC prediction for substations in frigid high-latitude regions using SSA-LS-SVM. *Smart Power* 48 (6), 54–60.
- Liu, S., Zhu, X., and Liu, L. (2020b). Fuzzy estimation model of LCC for 110 kV GIS substations based on fuzzy smoothing. *J. Guangxi Normal Univ. (Natural Science Edition)* 38 (5), 24–33. doi:10.16088/j.issn.1001-6600.2020.05.003
- Peng, X., Xiong, L., and Wen, J. (2016). Methods to improve short-term and ultra-short-term wind-power forecasting accuracy: a review. *Proc. CSEE* 36 (23), 6315–6326+6596. doi:10.13334/j.0258-8013.psee.161167
- Qiao, G., Guo, L., and Wu, Y. (2015). Substation LCC prediction model via GA-tuned LS-SVM. *China Electr. Power* 48 (11), 142–148.
- Shi, R., Zhang, L., and Li, Y. (2020). Intelligent statistical algorithm for equipment cost of power-grid projects based on physical ID. *Power Grid Clean Energy* 36 (2), 68–74.
- Su, H., Zhang, J., and Liang, Z. (2012). Substation life-cycle cost planning based on GIS spatial analysis and improved particle swarm optimization. *Proc. CSEE* 32 (16), 92–99. doi:10.13334/j.0258-8013.psee.2012.16.013
- Tao, Y., Chen, H., and Qin, X. (2018). Concepts, models and methods for short-term wind-power forecasting. *Electr. Power Eng. Technol.* 37 (5), 7–13. doi:10.19464/j.cnki.cn32-1541/tm.2018.05.002
- Wang, M., Zhang, Y., and Geng, P. (2020). Life-cycle cost prediction of substations based on optimized orthogonal partial least squares. *Smart Power* 48 (5), 119–124. doi:10.3969/j.issn.1673-7598.2020.05.019
- Wang, G., Dong, X., and Yang, F. (2021). Life-cycle cost modeling and sensitivity analysis for an unattended 110 kV substation in a coal mine. *China Min. Mag.* 30 (S1), 68–73.
- Wang, C., Yin, H., and Chen, S. (2022). Few-data wind-power forecasting considering spatial coupling. *South. Power Syst. Technol.* 16 (6), 75–81. doi:10.13648/j.cnki.issn1674-0629.2022.06.008
- Xiong, Z., Xiong, Y., and Xiong, Y. (2021a). Substation LCC prediction using quantum-behaved particle swarm optimization and LS-SVM. *Electr. Meas. Instrum.* 58 (6), 76–81. doi:10.19753/j.issn1001-1390.2021.06.011
- Xiong, Y., Zhan, Z., and Ke, F. (2021b). Prediction of substation maintenance and O&M cost using an improved BP neural network. *J. Electr. Power Sci. Technol.* 36 (4), 44–52. doi:10.19781/j.issn.1673-9140.2021.04.006
- Xue, Y., Wang, L., and Wang, S. (2019). Ultra-short-term wind-power prediction with a CNN-GRU hybrid model. *Renew. Energy* 37 (3), 456–462. doi:10.13941/j.cnki.21-1469/tk.2019.03.023
- Yang, L., Ying, L., and Wang, Y. (2017). Estimation of life-cycle cost for substation construction. *Comput. Simul.* 34 (1), 123–128.
- Yang, Z., Peng, X., and Xiong, Y. (2023). Short-term wind-power forecasting with neighboring wind-farm information and CNN-BiLSTM. *South. Power Syst. Technol.* 17 (2), 47–56. doi:10.13648/j.cnki.issn1674-0629.2023.02.006
- Ye, L., Li, Y., and Pei, M. (2023). Combination forecasting for wind power under cold-wave and small-sample conditions. *Proc. CSEE* 43 (2), 543–555. doi:10.13334/j.0258-8013.psee.221814
- Yin, H., Ou, Z., and Fu, J. (2021). A transfer-learning approach for wind-power prediction based on a serio-parallel deep-learning architecture. *Energy* 234, 121171. doi:10.1016/j.energy.2021.121271
- Zhang, Z. (2021). *Research on attention mechanism in neural networks*. University of Science and Technology of China. doi:10.27517/d.cnki.gzjku.2021.000623
- Zhang, C., Shao, Z., and Chen, F. (2022). Scenario data transfer for new-energy generation via conditional deep convolutional GAN. *Power Syst. Technol.* 46 (6), 2182–2190. doi:10.13335/j.1000-3673.pst.2021.1008
- Zhu, G., Zhu, Z., and Yuan, Y. (2024). Upper-bound assessment of GIS equipment failures using time-series models and deep learning. *J. Jilin Univ. (Engineering Technol. Edition)*, 1–8. doi:10.13229/j.cnki.jdxbgxb.20230307

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.