Received: 16 December 2023 | Revised: 9 August 2024 | Accepted: 3 September 2024

The Journal of Engineering

DOI: 10.1049/tje2.70009

The Institution of Engineering and Technology WILEY

LETTER

Electric vehicle charging load prediction based on graph attention networks and autoformer

Zeyang Tang[1] | Yibo Cui[1] | Qibiao Hu[2] | MinLiu Liu[1] | Wei Rao[1] | Xinshen Liu[2]

[1]State Grid Hubei Electric Power Research Institute, Wuhan, China

[2]School of Information Management Wuhan University, Wuhan, China

Correspondence
Qibiao Hu, School of Information Management Wuhan University, Wuhan, China.
Email: huqibiao@whu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 72074171; State Grid Hubei Electric Power Co., Ltd, Grant/Award Number: B3153221001D

Abstract

With the widespread popularity of electric vehicles in the domestic market, large-scale electric vehicle user data has been collected and stored. Highly accurate user-level charging load prediction has a wide range of application scenarios and great business value. However, most existing EV load prediction methods are modelled from the charging station perspective, ignoring the user's travel habits and charging demand. Therefore, this paper proposes a temporal spatial neural network based on graph attention and Autoformer to predict electric vehicle charging load. Firstly, the urban map of Wuhan is rasterized. Then, driving and charging data from the user level are aggregated into the raster module according to the time sequence, and a spatio-temporal graph data structure of user travel trajectory is constructed. Finally, the temporal spatial neural network is used to construct the EV charging load prediction model from the user's perspective. The experimental results show that, compared with other baseline prediction methods, the proposed method effectively improves the accuracy of the EV charging load prediction model by fully exploiting the distribution of EV user clusters in time and geographic space.

1 | INTRODUCTION

According to the statistics from the International Energy Agency, the transportation sector's energy consumption and its consequent carbon emissions constitute about 29% and 23% of the global total, respectively. This positions transportation as a major contributor to the ongoing energy crisis and environmental degradation. The shift towards carbon neutrality is driving the expansion of the electric vehicle market. However, this upsurge presents several challenges. Firstly, the issue of difficulty in charging, uneven distribution of charging stations, and the "information gap" between charging facilities often lead to a series of problems such as difficulty in finding charging stations and wastage of charging infrastructure. Secondly, the issue of distribution network load and the rapid expansion of the electric vehicle fleet significantly increases the load on the electrical grid [1, 2], resulting in grid overload, reduced power quality, and potential risks to the security and stability of the grid operation [3]. Unregulated charging can also lead to decreased power quality and circuit overload [4]. Therefore, the prediction of electric vehicle charging loads has vast potential in academia and significant commercial value, addressing the grid impact of widespread charging demand. Currently, both domestic and international academic research has explored various aspects of electric vehicle charging load prediction. This research typically focused on the following areas:

First, research from the perspective of charging stations: Arias et al. [5] introduced a model for predicting the power demand of charging stations, considering the spatiotemporal aspects, and forecasting the charging demand for different time periods. Second, research from the perspective of electric vehicle users: Different vehicle types often exhibit varying behavioural characteristics, hence, predictions of charging loads for each type can be made based on these distinct behaviours [6]. Additionally, considering the complexity of user charging decisions and the randomness of electric vehicle charging, predictions of charging station loads can also be facilitated through real-time interaction with multisource information [7]. Additionally, the Monte Carlo method has been used by researchers to simulate and predict the charging load of stations by replicating vehicle behaviour patterns [8, 9]. Wang et al. [10] calculated charging loads using a more stochastic momentary travel

probability model. Third, comprehensive research from multiple perspectives includes weather conditions and traffic flow as influential factors on user behaviour. Integrating the concept of travel chains with Markov decision process theory, researchers construct spatiotemporal transition models as effective methods for predicting charging loads [11]. Furthermore, Majidpour et al. [12] utilized customer profile data and socket measurement data to predict charging loads, employing techniques such as time-weighted dot products, sequence prediction algorithms, support vector regression, and random forest algorithms.

In recent years, the role of spatial factors in research has expanded. Li et al. [13] developed spatiotemporal distribution models for single-day electric vehicle charging load prediction using graph data and constraints from the road network. As machine learning and artificial intelligence have advanced, a wider array of methods has been applied in this field. Load sequences can be understood as a composite of waveforms at different frequencies. Wavelet transform is used to decompose these sequences, allowing for the independent prediction of each waveform segment, which aids in load forecasting [14]. Moreover, decision trees and random forest methods have been employed to create predictive models for electric vehicle charging demand based on historical traffic and weather data [15, 16]. Due to the complexity of constructing machine learning models and their lower generalization abilities, some scholars have started to use deep learning for charging load prediction. Artificial neural networks and long short-term memory networks can effectively predict the charging load for electric vehicles [17]. Aduam et al. [18] proposed a forecasting technique based on multi-feature data fusion to improve the accuracy of a deep learning model for EV charging station load forecasting. The proposed method uses multi-feature inputs based on historical weather data observations as multiple inputs to a long- and short-term memory model to achieve robust prediction of charging loads. Xin et al. [19] propose an EV charging load prediction method SC-CNN-LSTM based on spectral clustering and deep learning networks. Especially for the problem of insufficient amount of data on EV charging loads, we propose the use of Monte Carlo simulation to sample and simulation. This method solves the instability of the actual data. Graph convolution networks (GCNs) combine graph data structures with traditional convolution operations to create a new network structure [20], which can be used to learn and preserve spatial information. Yu et al. [21] proposed integrating graph convolutional neural networks and gated neural networks to capture both time and space information separately to complete load predictions. Based on the SSA-BPNN-MC model, Wang et al [22]. combined the Monte Carlo algorithm to simulate the initial charging time and initial power state of electric vehicles, calculated the charging duration and charging load of different types of electric vehicles, and obtained the regional total charging load curve. However, like the above work, the computational complexity problem and the data dependency problem are still not solved. Furthermore, the introduction of attention mechanisms has improved the accuracy of traffic flow prediction. Sheng et al. [23] improved predictive performance in spatiotemporal convolution networks by introducing attention mechanisms.

However, the above methods have high computational complexity and strong data dependence and are difficult to be extended to other influencing factors. What is more serious is that the existing charging load forecasting methods do not fully consider the spatial distribution of charging demand in the road network, which limits the accuracy and practicality of their forecasting. We show that EV charging demand has significant spatial distribution characteristics and has a significant impact on grid scheduling and charging infrastructure planning [24]. Therefore, this paper introduces a deep learning framework called GAT-autoformer, which leverages the graph attention network and Autoformer model to deeply learn the spatiotemporal characteristics of historical trajectories and charging data of electric vehicles. The model is trained and tested using real user driving data, aiming to accurately predict the user-centric charging load distribution.

2 | METHODS

2.1 | Graph

A graph consists of a set of given points and the connections between these points. The points represent entities and are commonly referred to as "vertices," while the connections between points represent the relationships between these entities and are typically referred to as "edges" or "arcs." Graphs can be mathematically represented as follows:

$$G = (V, E) \tag{1}$$

where G represents *a* graph. V represents the set of vertices (points) in graph G. E represents the set of edges (connections) in graph G. Specifically, each grid is regarded as an independent graph node. These nodes are connected by edges, where the edges represent the actual movement of electric vehicles between two adjacent grids. By defining such edges, the flow patterns and main driving paths of electric vehicles in urban space can be revealed. The structural diagram is shown in Figure 1.

2.2 | Graph attention network

The graph attention network (GAT) incorporates attention mechanisms with graph convolution networks [25]. It treats each node in the graph as a central node, calculates the similarity between each central node and its first-order neighbouring nodes, and uses the weighted sum of neighbouring node features to update the representation of that node.

The algorithmic process of the GAT involves calculating attention coefficients. For each node in the graph, features are extracted and then transformed through mapping to obtain vectors, which serve as inputs to a feedforward neural network. We can choose the *LeakyReLU* function as the activation function for the feedforward neural network. Afterwards, the output of the feedforward neural network is normalized to calculate the

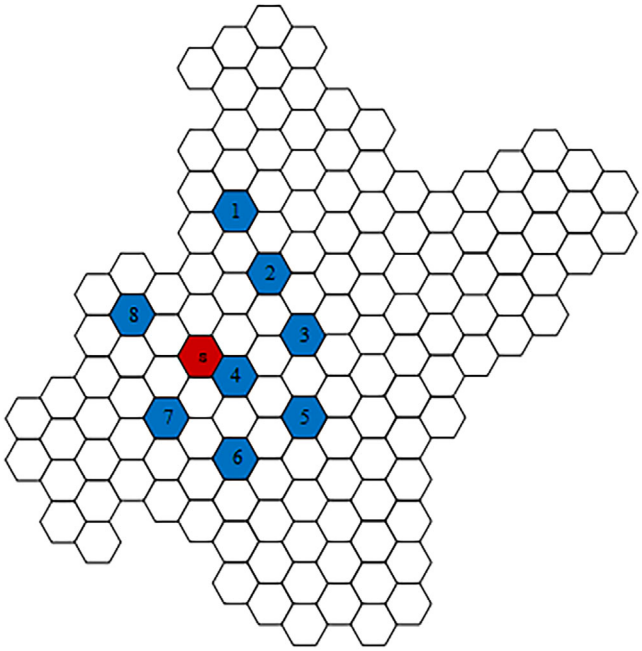

FIGURE 1 The schematic diagram of the gridded urban area of Wuhan.

attention coefficients. Finally, these attention coefficients can be applied to the prediction task, thereby enhancing predictive performance. This process can be represented by the following formula:

$$e_{ij} = LeakerRelu\left(\vec{\alpha}^T \left[\overrightarrow{Wh_i}\right] \left[\overrightarrow{Wh_j}\right]\right) \tag{2}$$

where e_{ij} represents the attention coefficient between node i and node j. $\vec{\alpha}^T$ is the weight vector for the attention coefficients. $\overrightarrow{Wh_i}$ and $\overrightarrow{Wh_j}$ are the results obtained by mapping the feature vectors of node i and node j, respectively.

To enhance the model's representational power, in the GAT, the output vectors of each attention head are concatenated and then averaged to obtain the input vector for each node in the entire space. This approach effectively utilizes graph information to capture relationships between entities, allowing for better modelling of associations between multiple entities. Additionally, by employing multi-head attention mechanisms, it is possible to capture relationships between multiple entities more effectively, thereby obtaining relationship information between entities. The calculation formula is as follows:

$$h_i = \Sigma_{k=1}^{K} \sigma\left(\Sigma_{j \in N_i} \alpha_{ij}^k W^k \overrightarrow{h_j}\right) \tag{3}$$

where h_i is the updated feature vector for node i after the attention mechanism. α_{ij}^k is the attention coefficient, representing the attention weight of node i on node j in the k-th attention head. W^k is the weight matrix for the k-th attention head. $\overrightarrow{h_j}$ is the original feature vector for node j. σ represents the activation function.

The GAT overcomes several GCN limitations concerning multi-layer graph representation, problem solving efficiency, and adaptability. Unlike GCN, which relies on complex computations with the Laplacian matrix, GAT updates node features by only considering first-order neighbouring nodes, eliminating the need for global graph processing. This feature allows GAT to handle different graph structures dynamically during training and testing. Furthermore, while GCN uniformly weighs neighbouring nodes in convolution operations, GAT utilizes an attention mechanism to apply varying weights to these neighbours. This grants GAT greater expressive power and a substantial advantage in performance over GCN.

2.3 | Time series model

Time series models are used to predict future information based on current or past limited information. Traditional time series models include Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Moving Average (ARMA), and Auto-Regressive Integrated Moving Average (ARIMA) models. In recent years, with the development of deep learning, deep forecasting models have made significant progress. Particularly, Transformer models [26], benefiting from self-attention mechanisms, have gained a substantial advantage in time series forecasting problems by capturing dependencies across time steps. However, there are still limitations in long-term sequence forecasting, as complex temporal patterns in long sequences make it challenging for attention mechanisms to discover reliable temporal dependencies. Transformer-based models often have to use sparse forms of attention mechanisms to address the issue of quadratic complexity, which can result in information utilization bottlenecks. Traditional Transformers and their variants typically use the following formula for calculating attention mechanisms:

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q represents the query vector, which measures the relevance of the query to the keys. K stands for the key vector, which represents the objects to be queried. V corresponds to the value vector, associated with the values for each key vector. d_k is the dimension of the key vectors. $SoftMax$ represents the $SoftMax$ function.

The issue with traditional Transformers is that they require similarity calculations with each element for each computation. For long input sequences, models are forced to use sparse dot products instead of individual dot products, which can lead to information loss. The Autoformer model addresses this by starting with sequence decomposition [27] and employing a decomposition architecture. This architecture excels at extracting predictive components from complex temporal patterns. Additionally, the Autoformer replaces the self-attention mechanism with an auto-correlation mechanism, which is grounded in stochastic process theory. Rather than calculating point-to-point correlations, it evaluates the correlations between sub-sequences, thereby efficiently capturing the relationships

between them. This method enhances sequence-level connection understanding, reduces computational complexity, and increases the efficiency of historical data utilization.

The Autoformer model consists of three main components: the sequence decomposition module, the auto-correlation mechanism, and the encoder and decoder.

2.3.1 | Sequence decomposition module

The sequence decomposition module of Autoformer decomposes the sequence into trend X_t and seasonal components X_s:

$$X_t = AvgPool\left(Padding\left(X\right)\right) \tag{5}$$

$$X_s = X - X_t \tag{6}$$

where X represents the original sequence. X_t represents the trend component of the sequence, representing long-term changes in the sequence. X_s represents the seasonal component of the sequence, representing short-term seasonal variations. *AvgPool* denotes the average pooling operation, which *Padding* ensures that the length of the sequence remains unchanged.

2.3.2 | Auto-correlation mechanism

The auto-correlation mechanism in Autoformer is designed to achieve efficient sequence-level connections to expand information utility. Typically, different cycles exhibit similar phase relationships, indicating the presence of similar sub-processes. Autoformer utilizes the inherent periodicity in sequences to design the auto-correlation mechanism. It calculates the sequence's auto-correlation coefficients to identify period-based dependencies. It then aggregates similar subsequences by shifting them in time based on the calculated period length, achieving time delay aggregation. The discovery of period-based dependencies is based on stochastic process theory, calculating correlation as a measure of confidence for unnormalized periodic estimates. Time delay aggregation aggregates information from similar subsequences based on the computed period length to achieve sequence-level connections.

First, by calculating the correlation $R_{XX}(\tau)$ between the original sequence and its lagged sequence, periodic similar subsequences are identified. The formula for this process is expressed as follows:

$$R_{XX}\left(\tau\right) = \lim_{L\to\infty} \frac{1}{L}\Sigma_{t=1}^{L} X_t X_{t-\tau} \tag{7}$$

where L represents the window length in the sequence, which corresponds to the number of samples selected from the sequence. τ represents the time interval.

Select the *Topk* highest $R(\tau)$, perform- Rolling on the original time series, convert the k $R(\tau)$ values into k probabilities using

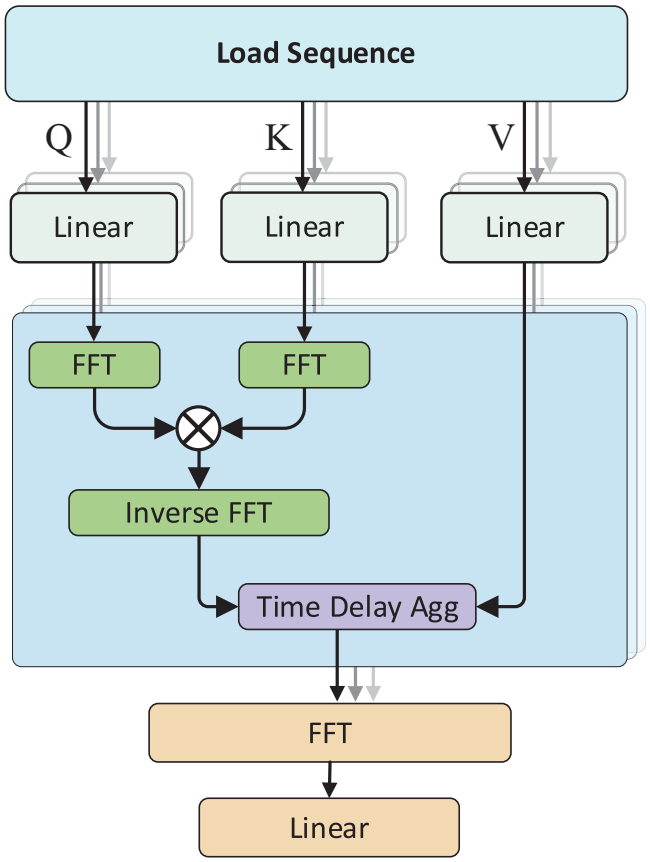

FIGURE 2 The structure of the autocorrelation mechanism.

Softmax, and then perform a weighted sum operation with their corresponding lagged time series. The formula is as follows:

$$\tau_1, \tau_2, \ldots, \tau_k = argTopk_{\tau\epsilon\{1,\ldots,L\}}\left(R_{Q,K}\left(\tau\right)\right) \tag{8}$$

$$\hat{R}_{Q,K}\left(\tau_1\right), \ldots, \hat{R}_{Q,K}\left(\tau_k\right) = SoftMax\left(\hat{R}_{Q,K}\left(\tau_1\right), \ldots, \hat{R}_{Q,K}\left(\tau_k\right)\right) \tag{9}$$

$$Auto - Correlation\left(Q, K, V\right) = \Sigma_{i=1}^{k}\, Roll\left(V, \tau_k\right)\hat{R}_{Q,K}\left(\tau_k\right) \tag{10}$$

where $argTopk(\cdot)$ is to get the arguments of the *Topk* autocorrelations and let $k = \lfloor c \times log\, L \rfloor$, c is a hyper-parameter. *SoftMax* is an activation function that scales numbers or logits into probabilities. The output of a *SoftMax* is a vector with probabilities of each possible outcome. $Roll\left(X, \tau\right)$ represents the operation to X with time delay τ, during which elements that are shifted beyond the first position are re-introduced at the last position. The structural diagram is depicted in Figure 2.

The Autoformer model addresses complex temporal patterns and information utilization bottlenecks through progressive decomposition and sequence-level connections, significantly improving long-term prediction performance. In the following sections of this article, the Autoformer model will also be used as a crucial component of the model.

2.4 | GAT-Autoformer

This paper introduces an electric vehicle charging load forecasting model based on GAT-Autoformer. As shown in Figure 1, the architecture of the electric vehicle charging load forecasting model based on GAT-Autoformer is mainly divided into two modules: the GAT block for handling spatial information and the Autoformer block for handling temporal information.

In the Autoformer block, the model processes temporal information through two components: the autoregressive block and the sequence decomposition block. The sequence decomposition block decomposes sequences of different time scales, in this study, daily, weekly, and monthly time scales. It calculates the correlations between different segments of time series using the multi-head attention mechanism in the auto-regressive block. The forward feature represents the current state of the data, while the backward feature represents the historical state of the data. Additionally, at each time step, there is a backward sequence from the previous moment to estimate the forward and backward sequences for the next moment. In the GAT block for handling spatial information, it also uses a graph convolutional structure with a multi-head attention mechanism. This allows multiple independent attention blocks to jointly learn spatial dependencies for different temporal states.

Therefore, the GAT-Autoformer model's spatiotemporal prediction combines modules for processing spatial and temporal information. This integration enables the simultaneous encoding of spatial and electric vehicle state information. The temporal prediction model then computes temporal dependencies within different spatial contexts. By taking in trajectory data from electric vehicle users at a specific scale, the model delivers the final predicted charging load results.

2.4.1 | Spatial modelling process

The GAT computes attention coefficients between nodes by employing an attention mechanism, which takes advantage of various dependency relationships between nodes. By incorporating these attention coefficients into an adjacency matrix, the resulting adjacency matrix is structured as follows:

The graph attention network computes attention coefficients between nodes by employing an attention mechanism, which takes advantage of various dependency relationships between nodes. By incorporating these attention coefficients into an adjacency matrix, the resulting adjacency matrix is structured as follows:

$$\tilde{\mathrm{A}} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1N} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N1} & \alpha_{N2} & \cdots & \alpha_{NN} \end{bmatrix} \tag{11}$$

where α represents the attention coefficient, and the product of the attention coefficients learned through the graph attention network and the hidden features represent the recalculated and updated node features. The introduction of the attention adjacency matrix allows for the representation of spatial nodes in the graph using different attention coefficients, making the spatiotemporal model proposed in this research more intuitive and dynamically representing spatial dependencies in the graph.

2.4.2 | Temporal modelling process

Autoformer is a novel time series forecasting model based on the Transformer architecture. It incorporates a deep decomposition framework, autoregressive mechanisms, and corresponding encoder-decoder components.

Time series decomposition refers to breaking down a time series into several components, each representing a certain type of underlying temporal pattern, such as periodic components and trend components (trend-cyclical). Sequence decomposition is integrated as an internal unit within Autoformer and embedded into the encoder-decoder architecture. During the prediction process, the model alternates between optimizing the predicted results and sequence decomposition. This means gradually separating trend and periodic components from the hidden variables, achieving a progressive decomposition. The sequence decomposition unit is based on the concept of sliding averages and undergoes smoothing to emphasize the trend component.

Due to the inherent uncertainty in predicting the future, it is common practice to decompose past sequences and then make individual predictions for each component. However, this approach can lead to predictions being constrained by the decomposition quality and overlooks the interactions between different components in the future.

The encoder takes an input sequence X_{en} of the past I time points' length, while the decoder takes inputs that include the seasonal component X_{des} and the periodic component X_{det}. The formula is represented as follows:

$$X_{ens}, X_{ent} = SeriesDecomp\left(X_{en\frac{I}{2}:I}\right) \tag{12}$$

$$X_{des} = Concat\ (X_{ens}, X_0) \tag{13}$$

$$X_{det} = Condat\ (X_{ent}, X_{Mean}) \tag{14}$$

in the equation, d represents the number of time series, I represents the length of the past time, and O, the future time length is split into two parts: the first $1/2$ part is obtained through decomposition X_{en}, and the latter O part is filled with 0s and the mean value "mean" to complete the sequence.

Assuming we have N encoding layers, the internal details of the l-th encoding layer $X_{en}^{l} = Encoder(X_{en}^{l-1})$ are as follows:

$$S_{en}^{l,1} = SD\left(AC\left(X_{en}^{l-1}\right) + \left(X_{en}^{l-1}\right)\right) \tag{15}$$

$$S_{en}^{l,2} = SeriesDecomp\left(FeedForward\left(S_{en}^{l,1}\right) + S_{en}^{l,1}\right) \tag{16}$$

where $S_{en}^{l,1}$ represents the output of the first attention sub-layer in the l-th encoding layer. SD is the operation for computing the standard deviation of the input. AC is the self-attention mechanism used to compute self-attention over the input. X_{en}^{l-1} represents the output of the $(l-1)$-th encoding layer.

Assuming there are M decoding layers, the internal details of the l-th decoding layer $X_{en}^{l} = Decoder(X_{en}^{l-1})$ are as follows:

$$S_{de}^{l,1}, T_{de}^{l,1} = SD\begin{pmatrix} AC\left(X_{de}^{l-1}\right) \\ AC\left(X_{de}^{l-1}\right) + X_{de}^{l-1} \end{pmatrix} \tag{17}$$

$$S_{de}^{l,2}, T_{de}^{l,2} = SD\left(AC\left(S_{de}^{l,1}, X_{en}^{N}\right) + S_{de}^{l,1}\right) \tag{18}$$

$$S_{de}^{l,3}, T_{de}^{l,3} = SD\left(FF\left(S_{de}^{l,2}\right) + S_{de}^{l,2}\right) \tag{19}$$

where $S_{de}^{l,1}$, $S_{de}^{l,2}$ and $S_{de}^{l,3}$ represent the outputs of the first, second, and third attention sub-layers in the l-th decoding layer, respectively, with residual connections. $T_{de}^{l,1}$, $T_{de}^{l,2}$ and $T_{de}^{l,3}$ are the residual connections of the first, second, and third attention sub-layers in the l-th decoding layer, respectively. X_{de}^{l-1} represents the output of the $(l-1)$-th decoding layer. FF denotes the feedforward neural network layer used for feedforward computations on the input.

The final output of the model is represented as:

$$W_s * X_{de}^{M} + T_{de}^{M} \tag{20}$$

where W_S is a matrix used for linear transformation of the input. X_{de}^{M} represents the output of the M-th decoding layer. T_{de}^{M} is the residual connection result for the M-th decoding layer.

Based on the progressive decomposition framework described above, the model can gradually break down hidden variables during the prediction process. It achieves this by utilizing autoregressive mechanisms and accumulating results to obtain predictions for the periodic and trend components separately. By combining spatial and temporal models, this paper establishes a novel deep self-decomposition model architecture based on both spatial and temporal information.

2.4.3 | Overall process of spatiotemporal fusion modelling

The overall process of user-level electric vehicle charging load forecasting based on the graph attention network and Autoformer model, as presented in this paper, is depicted in Figure 1. Here is a summary of the steps:

First, preprocessing of spatiotemporal trajectory data: The source data of electric vehicle user spatiotemporal trajectories are preprocessed to remove anomalies and missing values. Second, spatial grid generation: The map of Wuhan city is processed into a grid structure, dividing the Wuhan city area into rectangular grids with dimensions of 1 km by 1 km. Third, aggregation and spatiotemporal graph construction: Based on the gridded map, the preprocessed spatiotemporal trajectory data of electric vehicle users are aggregated to build a spatiotemporal graph data structure for a large cluster of electric vehicle users' travel trajectories. Then, GAT-Autoformer neural network training: Using the constructed spatiotemporal graph data, a GAT-Autoformer neural network is trained to uncover electric vehicle user charging patterns and predict the charging load for various city blocks in Wuhan during the next time period. Finally, load prediction aggregation: The predicted charging load values for each grid are aggregated to obtain the total forecasted electric vehicle charging load for the entire city of Wuhan.

TABLE 1 EV user charging dataset features.

Category	Feature	Feature description
User charging	Vin	Desensitized user vehicle ID
	start_time	Charging start time
	stop_time	Charging stop time
	Latitude	Current vehicle latitude (WGS84)
	charge_power	Charge power (%)
	start_soc	Charging start SOC (%)
	stop_soc	Charging stop SOC (%)
	Vehicleuse	Vehicle use
User trajectory	trip_id	Desensitized user vehicle ID
	collect_time	Data reporting time
	Speed	Current vehicle speed (km/s)
	Soc	Current vehicle remaining power (%)
	Longitude	Current vehicle longitude (WGS84)

3 | EXPERIMENTS

3.1 | Dataset

To validate the model's performance in this study, trajectory data and charging data from some electric vehicle users within the urban area of Wuhan, China, for the first quarter of 2022 were collected. Relevant user data concerning privacy has been anonymized, and coordinate points in the trajectory data were recorded at 5-minute intervals. The dataset comprises a total of 13 features categorized as user charging and user trajectory, as shown in Table 1. It includes three data types: integer, time, and string.

In this study, real user travel and charging data were used as experimental data, and the dataset was split in a 6:2:2 ratio, creating a training set, validation set, and test set. The spatial network connections in the dataset were established based on the actual road network.

3.2 | Performance evaluation metrics

According to the above data set division method, we adjusted the model's learning window to 3 days and the prediction window to 1 day, achieving a prediction accuracy of 93.25% on the

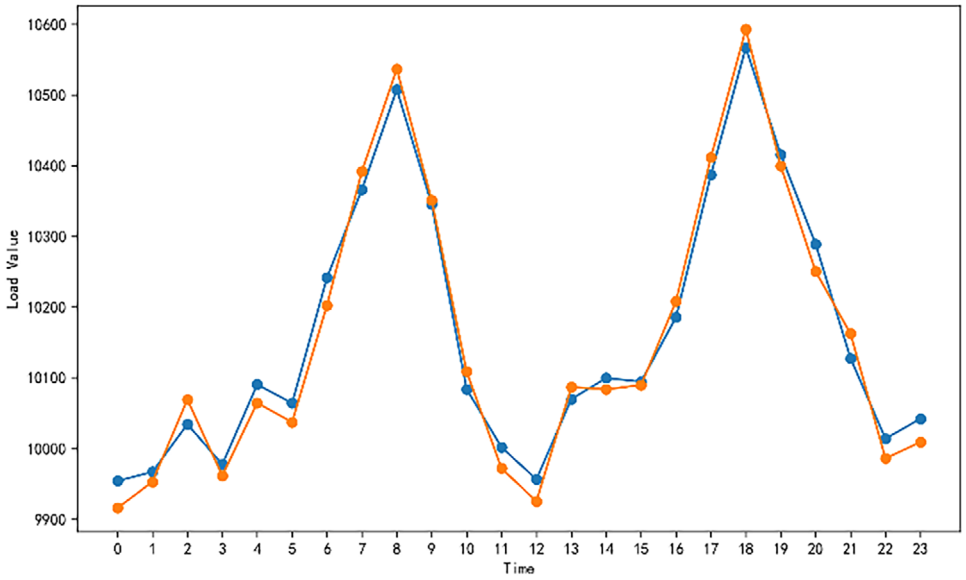

FIGURE 3 Comparison chart of load forecasting on the test set and actual values.

test set. Figure 3 is the result of visualizing an arbitrarily selected data point, which proves that our model can predict charging load quite well.

To evaluate the performance of the model, this paper utilizes the following performance evaluation metrics: mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). Their definitions are as follows:

$$MAE = \frac{1}{n}\Sigma_{t=1}^{n}\left|\tilde{y} - y\right| \tag{21}$$

$$MAPE = \frac{1}{n}\Sigma_{t=1}^{n}\left|\frac{\tilde{y} - y_t}{y_t}\right| \tag{22}$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{t=1}^{n}(\tilde{y}_t - y_t)^2} \tag{23}$$

where $\tilde{y}_t$ represents the model's predicted charging load. y_t represents the true charging load.

3.3 | Experimental results and analysis

3.3.1 | Model parameter configuration

All experiments were conducted on a Linux operating system (CPU: AMD 5950X, GPU: NVIDIA GeForce RTX 3090) using the PyTorch deep learning framework to train and predict with the GAT-Autoformer model. In this study, multiple-step predictions were made using a historical time window of $T = 12$ h. The performance of the model was evaluated using various time step lengths, including daily, weekly, and monthly, for charging demand load, and hyperparameters were set based on the validation set. In the model, all convolutional layers used 64 convolutional kernels. The batch size was set to 32, and training was performed using the Adam optimizer with an initial learning rate of 0.001. EarlyStopping was applied with a patience of 5.

TABLE 2 Comparison of the results of different prediction approaches.

Method of prediction	MAPE (%)	RMSE	MAE
LSTM	3.776	0.538	1.693
LSTM-attention	3.652	0.432	1.632
LSTnet	4.962	0.823	1.892
Informer	3.130	0.354	1.575
Autoformer	2.975	0.247	1.530
GAT-Autoformer	**2.687**	**0.211**	**1.479**

Abbreviations: GAT, graph attention network; LSTM, long short-term memory; MAE, mean absolute error; MAPE, mean absolute percentage error; RMSE, root mean square error.

3.3.2 | Experimental results analysis

To validate the superiority of the model, this paper compared the GAT-Autoformer model with five benchmark models. Table 2 presents the prediction results of different methods.

The baseline models used in this study are as follows:

1. Autoformer: An improved model based on the Transformer architecture for long sequence forecasting. It incorporates autoregressive mechanisms and a deep decomposition framework to enhance the prediction performance for long-term time series forecasting problems.
2. Informer: An improved model based on the Transformer architecture for long sequence forecasting. It modifies the structure of the Transformer to enhance the computation complexity of the self-attention mechanism and memory usage [28].

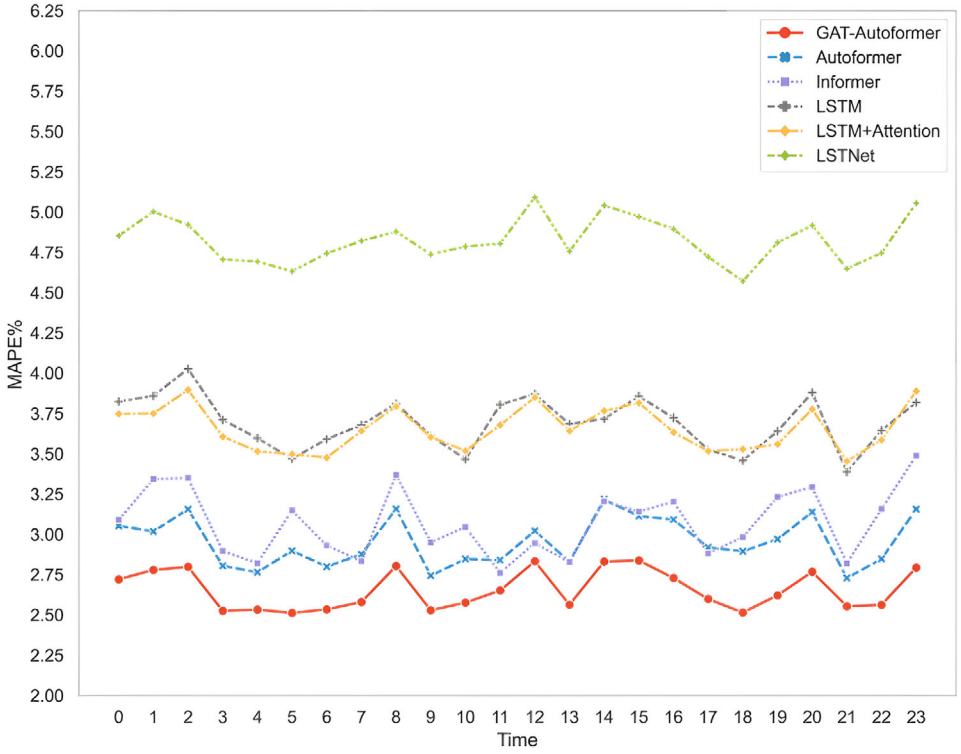

FIGURE 4 Graph attention network (GAT)-Autoformer based electric vehicle mean absolute percentage error results for charging load prediction. MAPE, mean absolute percentage error; LSTM, long short-term memory.

3. Long short-term memory (LSTM) network: Used to address the issues of vanishing gradients and exploding gradients during backpropagation in traditional RNN models and the lack of long-term memory [29].
4. LSTM-attention: It incorporates an attention mechanism into the traditional LSTM model, enabling it to better capture and predict correlations between time series. Identifying relationships between multiple time series can further enhance the model's predictive accuracy [30].
5. LSTNet: A model that combines convolutional neural networks (CNN) and recurrent neural networks (RNN) to extract short-term local dependency patterns between variables. LSTNet uses a CNN to extract short-term dependencies between variables and then passes them to an LSTM layer to capture long-term dependencies [31].

The table above presents the performance evaluation metric results for different prediction methods, with the GAT-Autoformer model achieving the best results. Comparing the results from Table 2, it is evident that the GAT-Autoformer model outperforms all the baseline prediction methods in terms of MAPE, MAE, and RMSE, indicating varying degrees of accuracy improvement across the three evaluation metrics. For instance, when compared to the Autoformer model, which performed the best among the baseline methods, the GAT-Autoformer model still manages to achieve a relative decrease of 0.288% in MAPE, a decrease of 3.3% in MAE, and a relative decrease of 14.6% in RMSE. Furthermore, as an example at a certain time, the load curve predictions from different prediction methods are shown in Figures 4–6. It is evident that the load curve predictions of the GAT-Autoformer model are closer to the actual daily load curve compared to the various baseline prediction methods, confirming the superior prediction accuracy of the GAT-Autoformer model.

To further illustrate the improvement in accuracy of the GAT-Autoformer model for electric vehicle charging load forecasting across different regions of Wuhan city, we can take the example of predicting the average charging load performance for 24 time periods in a day. As shown in Figures 3, 7, 8, with respect to the MAPE, MAE, and RMSE evaluation metrics, the GAT-Autoformer model demonstrates varying degrees of improvement in prediction accuracy for each time period compared to the various baseline prediction methods. This ensures an overall enhancement in the prediction accuracy of the charging load. It is evident that the GAT-Autoformer model exhibits a clear superiority in the prediction of urban charging load demand.

3.3.3 | Scalability analysis of the model

To validate the scalability of the GAT-Autoformer spatiotemporal prediction model, we applied it to traffic flow prediction tasks. Since traffic flow is closely related to spatial information, we believe that aggregating spatial information using GAT and applying attention weights to traffic flow prediction can

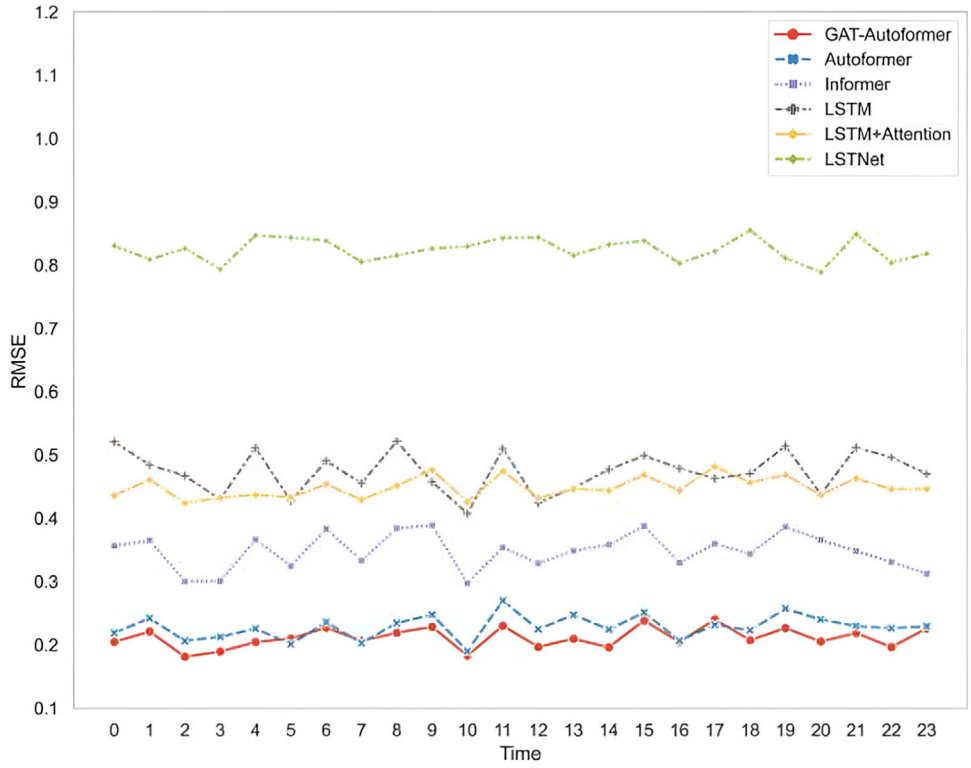

FIGURE 5 Graph attention network (GAT)-Autoformer based electric vehicle root mean square error results for charging load prediction. RMSE, root mean square error; LSTM, long short-term memory.

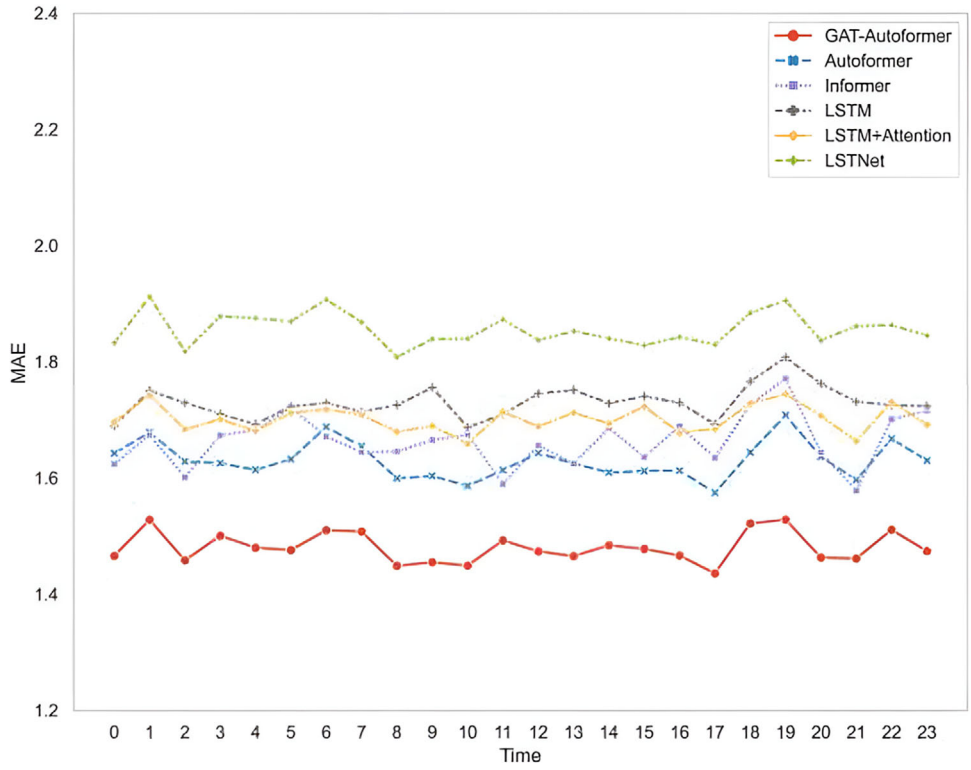

FIGURE 6 Graph attention network (GAT)-Autoformer based electric vehicle mean absolute error results for charging load prediction. MAE, mean absolute error; LSTM, long short-term memory.

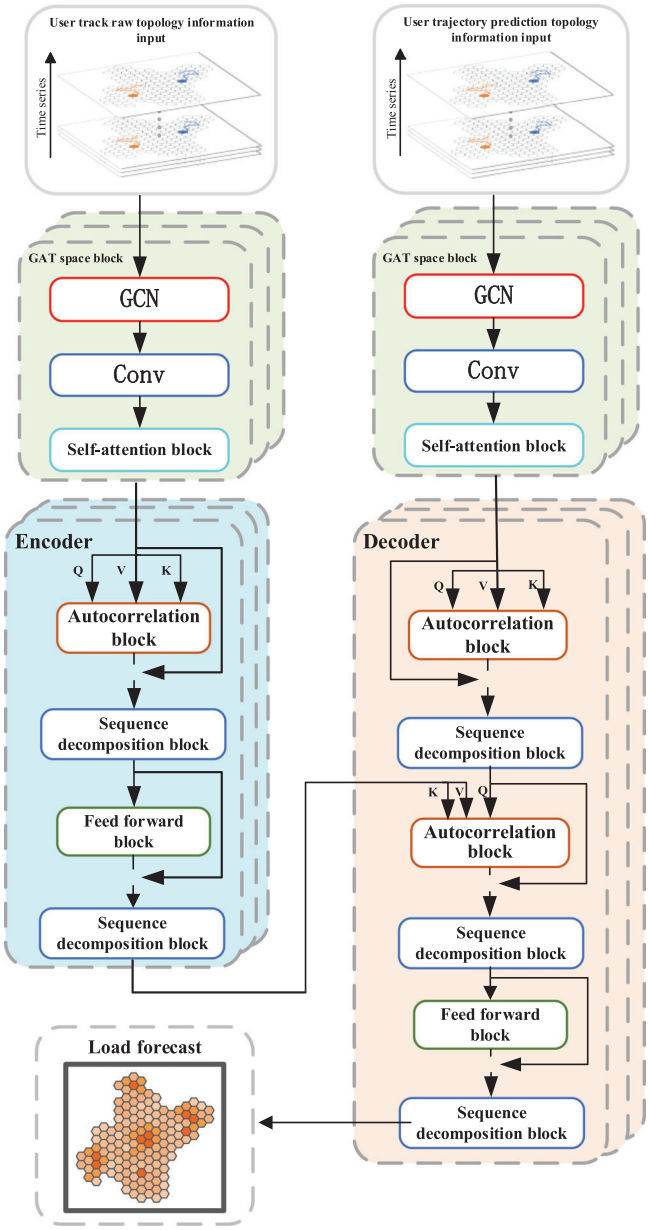

FIGURE 7 Model structure of load forecasting of the electric vehicle users based on graph attention network (GAT)-Autoformer. GCN, graph convolution networks.

effectively enhance prediction accuracy. Specifically, we chose the Wuhan region, rasterized it, and divided the grids into training, testing, and validation sets in a 6:2:2 ratio. We used the GAT to obtain attention weights for different grids and applied these weights to traffic flow sequence prediction. In this task, the learning window was set to 3 days and the prediction window to 1 day. Figure 9 shows the comparison between predicted and actual vehicle counts on the test set, and Table 3 presents the evaluation metrics for the test set.

The blue line in the above chart represents the actual number of vehicles, while the orange line indicates the predicted number of vehicles. As can be seen from the line chart, the model

TABLE 3 Comparison of the results of different prediction approaches.

Method of prediction	MAPE (%)	RMSE	MAE
LSTM	2.963	30.56	27.474
LSTM-attention	2.125	28.453	26.503
LSTnet	2.34	36.223	30.726
Informer	1.776	26.903	25.578
Autoformer	1.874	24.861	24.847
GAT-Autoformer	**1.060**	**19.87**	**16.245**

Abbreviations: GAT, graph attention network; LSTM, long short-term memory; MAE, mean absolute error; MAPE, mean absolute percentage error; RMSE, root mean square error.

can predict the actual number of vehicles with relatively high accuracy.

The aforementioned results indicate that the GAT-Autoformer has achieved the best performance in all evaluation metrics, thereby proving that the GAT-Autoformer model proposed in this study has good scalability and can be effectively applied to spatiotemporal sequence prediction tasks. The model is capable of capturing spatial features well and using them as auxiliary information to assist in the prediction of time series.

4 | CONCLUSION

This paper introduced a self-attentive temporal model, GAT-Autoformer, enhanced by graph attention, successfully applied to electric vehicle charging time series forecasting. GAT-Autoformer is a multi-layer neural network that includes graph attention enhancement and Autoformer, capturing dynamic spatiotemporal features in the travel patterns of electric vehicle users. The study employed real-world datasets of electric vehicle user travel trajectories and charging data from Wuhan for model training and validation. The results demonstrated that the proposed model outperforms other models in terms of electric vehicle load prediction accuracy. Moreover, the model showed significant computational efficiency improvements in the context of vast user travel trajectory data. This makes it suitable for use in charging load predictions with larger datasets and for real-time charging load demand forecasting.

Due to limitations in the experimental dataset, this paper constructed the input features for the proposed prediction model based on historical load sequences and geographical factors. In future research, considerations will be extended to include real-time traffic flow, points of interest (POI) labels, as well as the impact of meteorological factors and weather types on electric vehicle charging load prediction. Effective integration of meteorological and weather-type input features will be a focus. Additionally, traffic flow is influenced by various other external factors, such as the distribution of businesses in regions and societal events. In future work, these factors will be taken into account to increase the realism and controllable factors in the proposed approach, thereby improving its interpretability and predictive performance.

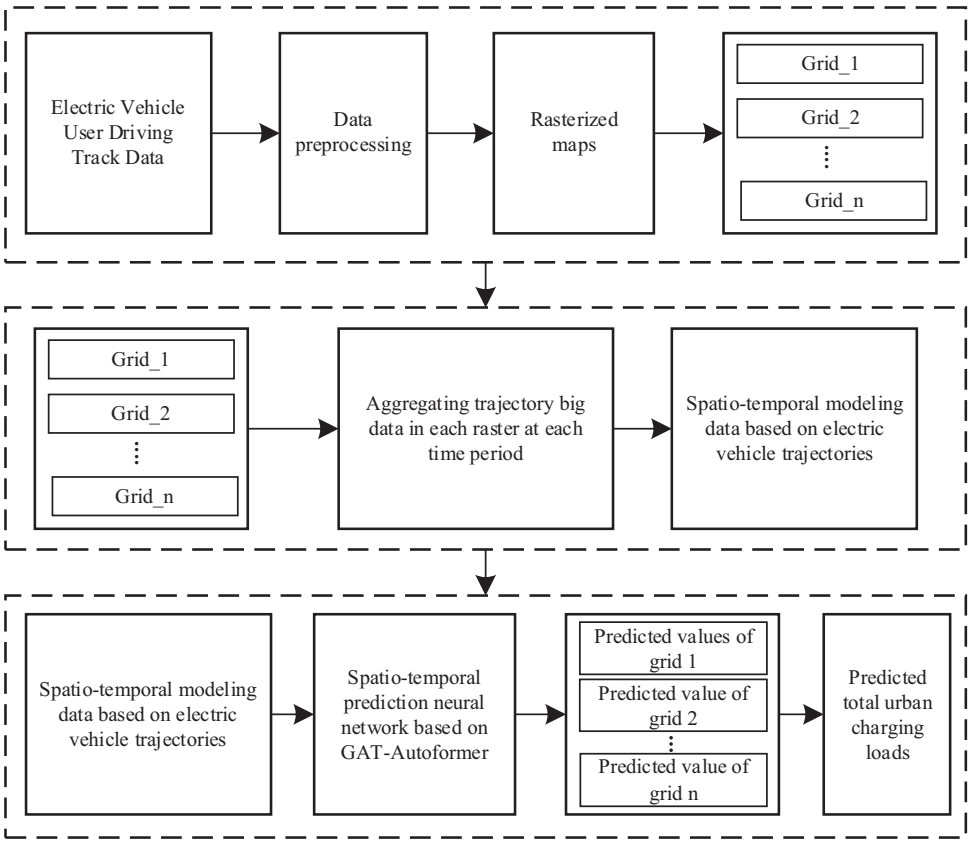

FIGURE 8 Flow chart of load forecasting of the electric vehicle users based on graph attention network (GAT)-Autoformer.

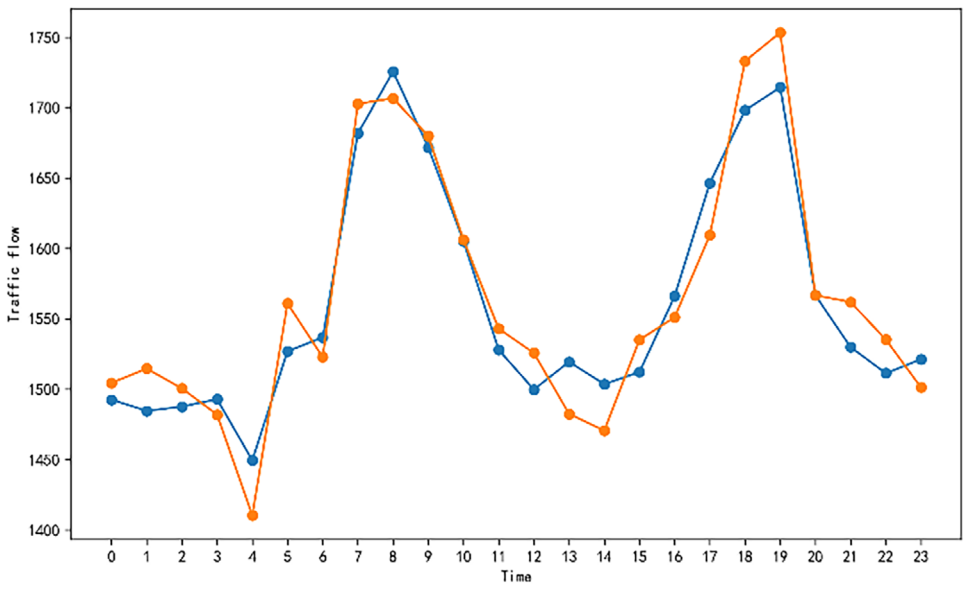

FIGURE 9 Comparison of 24 h traffic model predictions with real results.

AUTHOR CONTRIBUTIONS

Zeyang Tang conceptualized the study, contributed to visualization, and supervised the study. Yibo Cui proposed the methodology, validated the study, curated the data, wrote the original draft, and supervised the study. Qibiao Hu proposed the methodology, developed the software, conducted formal analysis, and curated the data. Minliu Liu contributed to software development and provided resources. Wei Rao reviewed

and edited the article, and administered the project. Xinshen Liu investigated the study. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

This work was funded by the National Natural Science Foundation of China (72074171), and the State Grid Hubei Electric Power Co., Ltd (B3153221001D).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are available from the corresponding author upon request.

ORCID

Qibiao Hu https://orcid.org/0009-0001-6328-8238

REFERENCES

1. Tang, W., Bi, S., Zhang, Y.J.A.: Online charging scheduling algorithms of electric vehicles in smart grid: an overview. IEEE Commun. Mag. 54(12), 76–83 (2016)
2. Niu, L., Zhang, P., Wang, X.: Hierarchical power control strategy on small-scale electric vehicle fast charging station. J. Cleaner Prod. 199, 1043–1049 (2018)
3. Nour, M., Chaves-Ávila, J.P., Magdy, G., et al.: Review of positive and negative impacts of electric vehicles charging on electric power systems. Energies 13(18), 4675 (2020)
4. Lopes, J.A.P., Soares, F.J., Almeida, P.M.R.: Integration of electric vehicles in the electric power system. Proc. IEEE 99(1), 168–183 (2010)
5. Arias, M.B., Kim, M., Bae, S.: Prediction of electric vehicle charging-power demand in realistic urban traffic networks. Appl. Energy 195, 738–753 (2017)
6. Ge, X., Shi, L., Fu, Y., et al.: Data-driven spatial-temporal prediction of electric vehicle load profile considering charging behavior. Electr. Power Syst. Res. 187, 106469 (2020)
7. Zhang, M., Sun, Q., Yang, X.: Electric vehicle charging load prediction considering multi-source information real-time interaction and user regret psychology. Power Syst. Technol. 46(02):632–645 (2022)
8. Tian, J., Lv, Y., Zhao, Q., et al.: Electric vehicle charging load prediction considering the orderly charging. Energy Rep. 8, 124–134 (2022)
9. Fan, L., Chen, L., Luo, W., et al.: Spatial and temporal distribution model of electric vehicle load considering different urban functional areas. Electr. Power Constr. 42(06), 67–75 (2021)
10. Wang, H., Zhang, Y., Mao, H.: Charging load forecasting method based on instantaneous charging probability for electric vehicles. Electr. Power Autom. Equip. 39(03), 207–213 (2019)
11. Cheng, S., Wei, Z., Shang, D., et al.: Charging load prediction and distribution network reliability evaluation considering electric vehicles' spatial-temporal transfer randomness. IEEE Access 8, 124084–124096 (2020)
12. Majidpour, M., Qiu, C., Chu, P., et al.: Forecasting the EV charging load based on customer profile or station measurement? Appl. Energy 163, 134–141 (2016)
13. Li, H., Du, Z., Chen, L., et al.: Trip simulation based charging load forecasting model and vehicle-to-grid evaluation of electric vehicles. Autom. Electr. Power Syst. 43(21), 88–96 (2019)
14. Zhang, J., Zhang, Q., Ma, Y.: Short-term load frequency domain prediction method based on improved random forest and density-based spatial clustering of applications with noise. Control Theory Appl. 37(10), 2257–2265 (2020)
15. Deb, S., Gao, X.Z.: Prediction of charging demand of electric city buses of Helsinki, Finland by random forest. Energies 15(10), 3679 (2022)
16. Arias, M.B., Bae, S.: Electric vehicle charging demand forecasting model based on big data technologies. Appl. Energy 183, 327–339 (2016)
17. Zhu, J., Yang, Z., Chang, Y., et al.: A novel LSTM based deep learning approach for multi-time scale electric vehicles charging load prediction. In: Proceedings of the 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia), pp. 3531–3536. IEEE, Piscataway, NJ (2019)
18. Aduama, P., Zhang, Z., Al-Sumaiti, A.S.: Multi-feature data fusion-based load forecasting of electric vehicle charging stations using a deep learning model. Energies 16(3), 1309 (2023)
19. Xin, F., Yang, X., Wang, B., Xu, R., Mei, F., Zheng, J.: Research on electric vehicle charging load prediction method based on spectral clustering and deep learning network. Front. Energy Res. 12, 1294453 (2024)
20. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv:1609.02907 (2016)
21. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv:1709.04875 (2017)
22. Wang, Q., Yang, X., Yu, X., Yun, J., Zhang, J.: Electric vehicle participation in regional grid demand response: potential analysis model and architecture planning. Sustainability 15(3), 2763 (2023)
23. Sheng, N., You, F., et al.: Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. Proc. AAAI Conf. Artif. Intell. 33(01), 922–929 (2019)
24. Paevere, P., Higgins, A., Ren, Z., Horn, M., Grozev, G., McNamara, C.: Spatio-temporal modelling of electric vehicle charging demand and impacts on peak household electrical load. Sustainability Sci. 9, 61–76 (2014)
25. Velikovi, P., Cucurull, G., Casanova, A., et al.: Graph attention networks. arXiv:1710.10903 (2017)
26. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. Curran Associates Inc., Red Hook, NY (2017)
27. Wu, H., Xu, J., Wang, J., et al.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. arXiv.2106.13008 (2021)
28. Zhou, H., Zhang, S., Peng, J., et al.: Informer: Beyond efficient transformer for long sequence time-series forecasting. Proc. AAAI Conf. Artif. Intell. 35(12), 11106–11115 (2021)
29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997)
30. Wang, Y., Huang, M., Zhu, X., et al.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 606–615. Association for Computational Linguistics, Stroudsburg, PA (2016)
31. Lai, G., Chang, W.C., Yang, Y., et al.: Modeling long-and short-term temporal patterns with deep neural networks. In: Proceedings of the the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 95–104. Association for Computing Machinery, New York, NY (2018)
32. Tang, Z., Cui, Y., Hu, Q., et al.: Electric vehicle charging load prediction based on graph attention networks and autoformer. Authorea. https://doi.org/10.22541/au.169027924.44245143/v1

How to cite this article: Tang, Z., Cui, Y., Hu, Q., Liu, M.L., Rao, W., Liu, X.: Electric vehicle charging load prediction based on graph attention networks and autoformer. J. Eng. 2024, e70009 (2024). https://doi.org/10.1049/tje2.70009