



ARTICLE

A MCG-GFAM-MRDCM Model for Accurate Building Electricity Load Forecasting

Chuan Lin^{*}, Weixian Chen and Guangtao Hao^{*}

College of Intelligent Manufacturing, Putian University, Putian, China

^{*}Corresponding Authors: Chuan Lin. Email: linchuan03@ptu.edu.cn; Guangtao Hao. Email: sduhgt@126.com

Received: 12 February 2026; Accepted: 08 April 2026; Published: 12 July 2026

ABSTRACT: Accurate building electricity load forecasting (BELF) can provide a regulatory basis for building energy management systems and promote the transition of buildings toward low-carbon and intelligent operation modes. However, building electricity load is influenced by historical loads, as well as outside environmental conditions such as humidity and temperature, which reduces the prediction accuracy of models. To tackle these challenges, this study presents a BELF model, which consists of a modal component grouping approach, grouped feature attention mechanism, and multi-scale residual depthwise convolution memory module. First, the modal component grouping method analyzes building electricity load in the time domain, frequency domain (via fast fourier transform, FFT), and complexity (via sample entropy, SE), and then performs clustering to achieve precise decomposition of load components with different fluctuation characteristics. Second, the grouped feature attention mechanism assigns suitable importance to various input features to emphasize key factors affecting prediction accuracy. Third, the multi-scale residual depthwise convolution memory module mitigates the impact of long and short-term load variations on BELF by employing residual blocks of depthwise convolution layers with different kernel sizes. Meanwhile, gated recurrent units are used to identify the time-dependent trends of building load. Experimental results on public buildings show that the proposed model outperforms existing models, achieving more than 2.4% improvement in MAPE prediction performance.

KEYWORDS: Building electricity load forecasting; modal component grouping method; depthwise convolutional neural network; attention mechanism

1 Introduction

Against the backdrop of the vigorous development of the global economy and the continuous advancement of urbanization, the industrial scale of the construction industry has been expanding [1,2]. The growing requirements for the comfort of living and working environments have directly led to an increase in building electricity energy demand [3,4]. At the same time, the increasingly scarce traditional energy resources and environmental problems caused by carbon emissions, such as global warming and frequent extreme weather, have brought enormous challenges to human sustainable development [5,6]. Against this background, how to efficiently manage building energy and reduce energy consumption has become a key issue to be solved [7,8]. Building electricity load forecasting (BELF) technology can provide support for the rational allocation and efficient utilization of building energy [9].

In recent decades, the field of BELF modeling has rapid development [10]. Currently, the main methods for BELF comprise statistics-based learning methods, machine learning methods, and deep

learning methods [11]. Statistical learning methods are capable of leveraging historical building load data directly to autonomously identify the underlying patterns contained in the data [12,13]. Ref. [14] utilized the autoregressive integrated moving average (ARIMA) model as a fundamental statistical approach to address building energy forecasting problems. Ref. [15] employed regression analysis to predict the linear relationships and consumption trends of residential building loads. While statistical learning approaches are straightforward to implement, they typically demand that data possess stationary properties. Nevertheless, actual building load does not meet this requirement, which reduces the precision of BELF.

Numerous methods grounded in machine learning have been developed by researchers to overcome the limitations of statistical learning and to identify the nonlinear relationships characteristic of building electricity consumption. Ref. [16] employed a support vector regression (SVR) method with historical temperature and load as inputs to forecast short-term building electricity demand. Ref. [17] adopted random forest (RF) to perform replacement sampling from the original data, constructing multiple sub-datasets and training decision trees respectively to identify the intricate features of load data. Secondly, load data of different power consumption activities were separately constructed into matrices to extract the unique change patterns of load types. In [18], the k-means method was initially utilized to cluster the load data of residential buildings, aiming to identify load data sets that share similar characteristics. Then, a neural network was used to extract complex load patterns for BELF. While machine learning methods can achieve more accurate forecasting results compared to statistical learning methods, they also have drawbacks. Conventional machine learning approaches usually depend on manually engineered features to identify and choose attributes relevant to BELF [9]. It is very challenging to manually design effective features, which requires domain expert knowledge and limits the performance of BELF [10].

Over recent years, deep learning methods have found widespread use in BELF research, given that they can adeptly retrieve the nonlinear attributes of historical loads to achieve accurate BELF predictions [11,12]. We classify the existing deep learning methods into three aspects: data processing methods, feature extraction methods, and attention mechanisms [19–30]. In terms of data processing methods, Ref. [19] designed a forecasting model based on empirical mode decomposition (EMD) and LSTM. First, EMD was used to decompose the input load into multiple components, so as to reduce the non-stationarity of the original load. Second, LSTM learned each component separately and output the forecasting results. Ref. [20] designed a prediction method which consists of variational mode decomposition (VMD) and temporal convolutional network (TCN). This model was designed to extract multi-scale load frequency features and model the temporal dependence relationships of loads. Ref. [21] designed a prediction model based on complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and long short-term memory network (LSTM). Firstly, CEEMDAN was used to process the nonlinear features of the input load. Secondly, LSTM was introduced to capture the temporal patterns of the load, so as to output the predicted load. Ref. [22] first decomposed the historical load using CEEMDAN to obtain components with different frequencies. Then, sample entropy (SE) was introduced to reconstruct each component. Finally, the LSTM method was adopted to separately predict the reconstructed component sequences, and the final load prediction result was obtained. In terms of feature extraction, Ref. [23] developed a prediction method based on CNN and SVR to mine the nonlinear characteristics between load and temperature, thereby realizing the hourly electricity load prediction for residential buildings. Ref. [24] adopted long short-term memory (LSTM) network to identify the periodic characteristics of building load, so as to achieve accurate load prediction. Ref. [25] developed a hybrid model combining CNN and LSTM to forecast peak loads. The model first employed a CNN to identify local load features from historical load and temperature data. Following this step, the Long Short-Term Memory (LSTM) network was applied to capture long-range temporal correlations present in the sequences. However, the complexity of its gating architecture

means that processing large datasets often involves significant computational costs and extended training periods, thereby limiting the model's generalizability. Ref. [26] proposed a deep learning method comprising LSTM and GRU networks to generate 1-h-ahead electricity load forecasts for residential buildings. Ref. [27] proposed a CNN model based on residual network (ResNet) to predict building loads one hour ahead. The model first employed a CNN to capture local features in the data, and then utilized the ResNet architecture to alleviate the vanishing gradient issue and improve learning efficiency of model. Ref. [28] proposed a forecasting framework which consists of CNN, GRU, and ResNet. The forecasting model first identified the spatial features and temporal patterns of building load through CNN and GRU, respectively. Second, the fused features were used as the input of ResNet to improve the forecasting accuracy. In terms of the attention mechanism, Ref. [29] developed a method based on the dual attention mechanism to allocate input features, thereby improving the prediction accuracy of the model. Ref. [30] proposed a forecasting model that integrates CEEMDAN, CNN, LSTM and self-attention (SA) to achieve accurate household electricity load forecasting. Furthermore, it is important to recognize that building electricity load prediction does not exist in isolation but serves as a crucial foundation for downstream dynamic energy management. Under non-stationary and dynamic operating conditions, advanced control strategies such as Model Predictive Control (MPC) and adaptive frameworks are highly required to ensure system stability and economic efficiency. For instance, recent studies have demonstrated the immense potential of adaptive mechanisms, such as adaptive event-triggered tracking control via switching functions [31], in handling dynamic disturbances. Similarly, advanced MPC frameworks, such as intrusion-detector-dependent distributed economic MPC for load frequency regulation [32], rely heavily on real-time and accurate forward-looking data to optimize energy dispatch under complex cyber-physical environments. However, the performance and reliability of these adaptive and MPC-based control frameworks fundamentally depend on the accuracy of the underlying load forecasting models. Therefore, developing a robust forecasting model that can effectively capture the non-stationary and dynamic characteristics of electricity loads—which is the primary focus of this study—is an indispensable prerequisite for enabling these advanced dynamic control applications.

Although deep learning methods for building load forecasting achieve higher prediction accuracy than machine learning and statistical learning methods, they still face certain limitations:

- (1) Existing methods usually adopt decomposition techniques to capture the nonlinear relationships of input data. However, they generally ignore the mining of frequency-domain information of load signals and lack quantitative analysis of their complex characteristics. This makes it difficult for the models to accurately capture the load features of different fluctuation types, ultimately limiting the accuracy of BELF.
- (2) Existing building electricity load feature extraction methods based on convolutional neural network (CNN) overlook the dynamic characteristics of building electricity load data and the differences between long-term and short-term fluctuations. On the one hand, the fixed receptive field is difficult to adapt to the dynamic characteristics of the data, resulting in insufficient accuracy in extracting key electricity consumption features. On the other hand, single-scale convolution struggles to address the differences between long-term and short-term load fluctuations, which further limits the efficiency of feature extraction.
- (3) The aforementioned attention mechanisms usually ignore the different importance and inherent correlations among features. As a result, when multiple types of features act together, the model cannot effectively capture the unique impacts of each group of features, ultimately leading to a decline in prediction accuracy.

To address the aforementioned issues, this paper proposes a novel MCG-GFAM-MRDCM prediction model. The main contributions of this paper are as follows:

- (1) This paper designs a modal component grouping method to explore the frequency-domain information and complex characteristics of electrical load, thereby alleviating the problem that traditional decomposition methods ignore such information and thus limit the accuracy of BELF. Firstly, ensemble empirical mode decomposition (EEMD) is applied to conduct time-domain analysis on building electrical load signals, yielding stable intrinsic mode function (IMF) components and a final residual trend term. Secondly, fast fourier transform (FFT) is used to analyze the frequency-domain characteristics of each IMF component and extract the dominant frequency of each IMF component. Thirdly, sample entropy (SE) is employed to calculate the complexity of each IMF component. Finally, the K-means clustering algorithm is adopted to classify the IMF components into weak components, medium components, and strong components.
- (2) This paper designs a multi-scale residual depthwise convolutional memory module (MRDCM) based on depthwise convolution and gated recurrent unit (GRU). By virtue of multi-scale depthwise convolution, MRDCM can automatically configure independent convolution kernels to enhance the ability of extracting long-term and short-term load features. Subsequently, the GRU is capable of capturing the temporal patterns in load, thus enhancing the accuracy of BELF.
- (3) This paper presents a grouped feature attention mechanism (GFAM) to allocate adaptive weights to each exogenous variable. It helps the model automatically identify the impact of key variables on BELF, thereby boosting the model's training efficacy.

The rest of this research is structured as follows: [Section 2](#) elaborates on the proposed MCG-GFAM-MRDCM prediction model and its component modules in detail. [Section 3](#) evaluates the performance of various modules and contrasts the prediction outcomes of the MCG-GFAM-MRDCM with those of existing models. [Section 4](#) summarizes the research content of this research.

2 Description of Building Electricity Load Forecasting

[Fig. 1](#) depicts the overall working process of this research. First, preprocessing operations are performed on the original building-related data, mainly including missing data imputation and abnormal data correction, to ensure the integrity and reliability of the input data and lay the foundation for subsequent analysis. Second, the ensemble empirical mode decomposition (EEMD) method is adopted to decompose the building electrical load data, resulting in a series of modal components with various frequencies. Subsequently, the frequency characteristics of each modal component are calculated via fast fourier transform (FFT), the complexity of the modal components is quantitatively analyzed using sample entropy (SE), and the decomposed modal components are classified with the help of the K-means clustering algorithm. On this basis, the preprocessed data, building load data, meteorological data (including parameters such as humidity, temperature, and pressure) and calendar data (such as time information and holiday indicators) are used as model input variables and imported into the constructed prediction model. This prediction model integrates a grouped feature attention mechanism and a multi-scale residual depthwise convolutional memory module. During the model training, validation, and testing phases, the model can adaptively focus on important load components and external influencing variables. In this process, the method can effectively extract the long-term and short-term variation characteristics of building load, thereby accurately capturing its temporal evolution pattern. Finally, the building electrical load prediction results are generated, and three core performance metrics, namely mean absolute percentage error (MAPE), root mean square error (RMSE), and mean absolute error (MAE), are used to conduct a quantitative assessment of the prediction error.

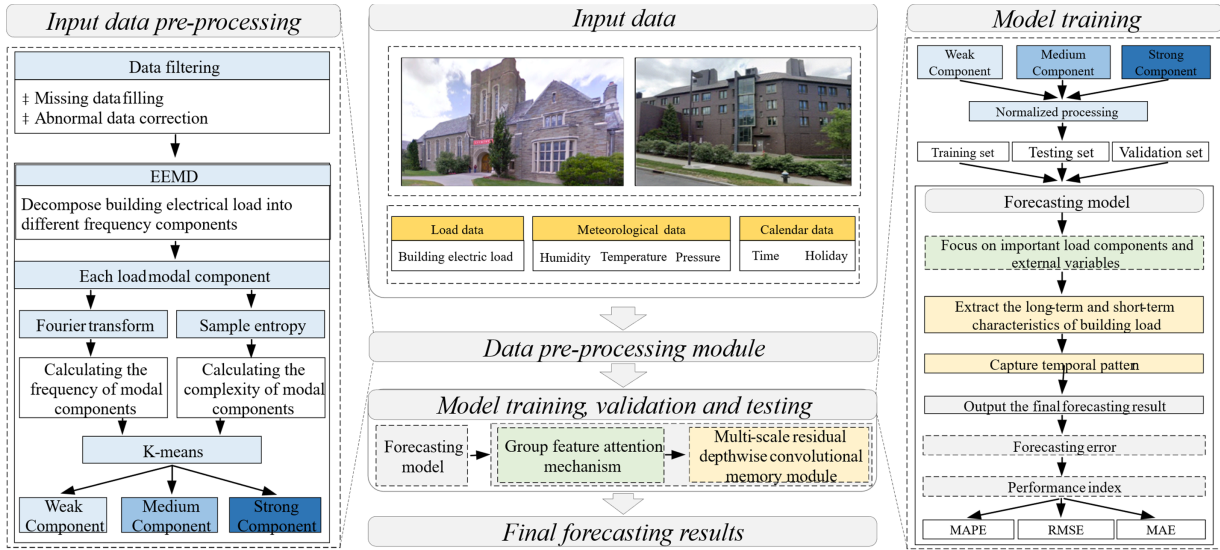


Figure 1: Flow chart of this paper.

3 MCG-GFAM-MRDCM Method

The structure of MCG-GFAM-MRDCM is shown in Fig. 2, which includes the modal component grouping method (MCG), grouped feature attention mechanism (GFAM), and multi-scale residual depthwise convolution memory module (MRDCM).

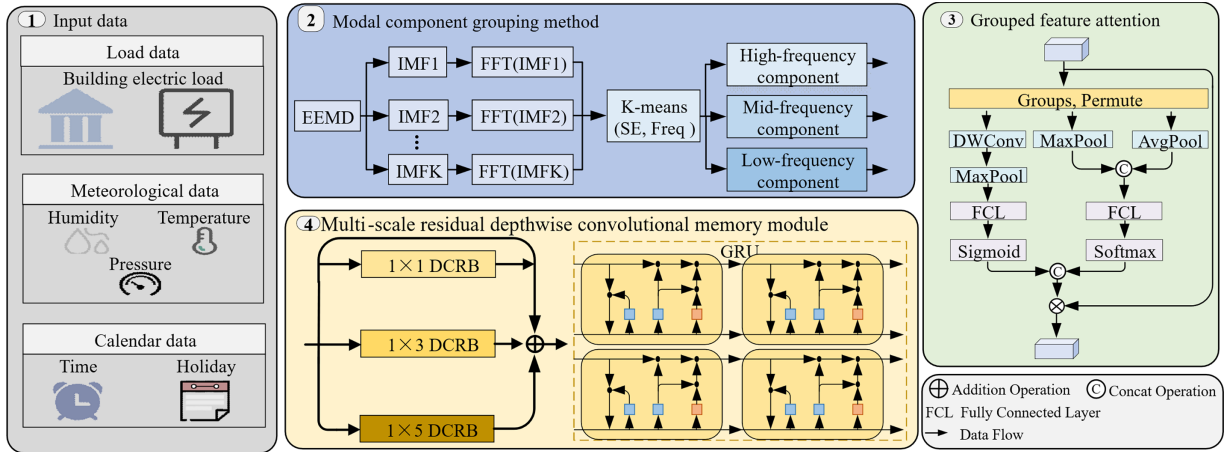


Figure 2: The structure of the MCG-GFAM-MRDCM method.

3.1 Modal Component Grouping Method

Traditional decomposition methods often focus only on the time-domain characteristics of electrical loads while generally neglecting the exploration of frequency-domain information and the complex characteristics of electrical loads. However, relying solely on the time domain limits the model's ability to distinguish between regular periodic loads and random noise. Therefore, this paper designs a modal component grouping method across the time domain, frequency domain, and complexity, as presented in Fig. 2. Specifically, the time-domain decomposition provides the foundational separation of raw signals; the frequency-domain analysis identifies the periodic behaviors of building energy use; and the complexity analysis quantifies the

randomness caused by sudden human activities. By integrating these three dimensions, the proposed method can map mathematical IMFs to actual physical building load components, significantly reducing the learning difficulty for subsequent prediction models.

3.1.1 Time-Domain Characteristic Analysis

Building electrical loads typically exhibit complex nonlinear characteristics. Using them directly for forecasting will reduce the forecasting accuracy of the method. This paper adopts EEMD to decompose complex load signals, thereby improving the accuracy of BELF. Fig. 3 presents the decomposition results of EEMD. The specific implementation steps are as follows:

Step 1: EEMD constructs a noisy building electrical load signal. Given the building electrical load data as $x(t)$ and the Gaussian white noise added in the i -th time as $\omega_i(t)$, the building electrical load data with noise added in the i -th group is formulated as:

$$x_i(t) = x(t) + \delta \cdot \omega_i(t) \quad (1)$$

where δ denotes the noise standard deviation, N stands for the total number of noise additions, and i denotes the serial number of an individual noise addition.

Step 2: For $x_i(t)$ obtained in the previous step, execute the empirical mode decomposition (EMD) process respectively. The expression is as follows:

$$x_i(t) = \sum_{k=1}^K IMF_{i,k}(t) + r_i(t) \quad (2)$$

where $IMF_{i,k}(t)$ is the k -th IMF load component decomposed from the i -th group of noisy load signals, $r_i(t)$ is the final residual load component after decomposing the i -th noisy signal, and K is the total number of IMFs obtained from a single EMD decomposition.

Step 3: Perform an ensemble average on all IMFs to produce the final IMFs of the EEMD decomposition. Meanwhile, EEMD averages the N groups of residual components to output the final residual trend term. The expressions are as follows:

$$IMF_k(t) = \frac{1}{N} \sum_{i=1}^N IMF_{i,k}(t) \quad (3)$$

$$r(t) = \frac{1}{N} \sum_{i=1}^N r_i(t) \quad (4)$$

where $IMF_k(t)$ is the k -th stable IMF component obtained by decomposing the input load signal $x(t)$, $r(t)$ is the final residual component after decomposition, and $r_i(t)$ is the residual component after decomposing the i -th group of noisy signals.

Step 4: Repeat Steps 2–3 until the ensemble average calculation is completed for all IMFs. Finally, the input load $x(t)$ is the sum of all final IMFs and the average residual component, which is expressed as follows:

$$x(t) = \sum_{k=1}^K IMF_k(t) + r(t) \quad (5)$$

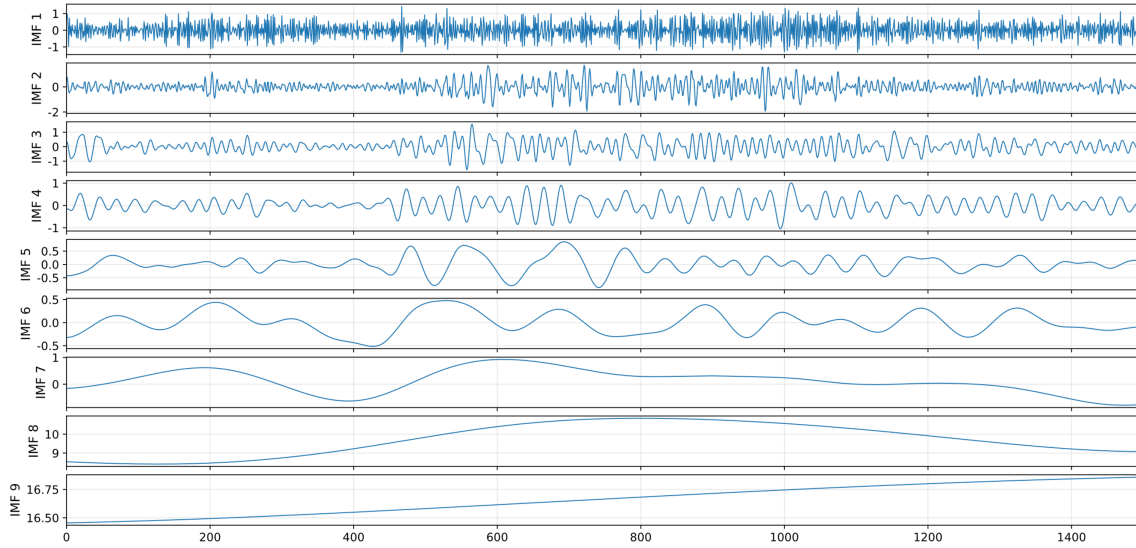


Figure 3: Decomposed results of electricity load data using EEMD.

3.1.2 Frequency-Domain Characteristic Analysis

To accurately extract the frequency characteristics of the signal, this paper uses fast fourier transform (FFT) to calculate the main frequency of each IMF decomposed by EEMD, as shown in Fig. 4. The specific steps are as follows:

Step 1: Apply FFT to each IMF component decomposed by EEMD for frequency-domain transformation to obtain the frequency spectrum. The expression is as follows:

$$F(f) = \int_{-\infty}^{\infty} IMF_k(t) \cdot e^{-j2\pi ft} dt \quad (6)$$

where N is the length of the time series, f is the frequency, and T is the time interval.

Step 2: To obtain the amplitude spectrum $A(f)$ of each IMF component, combine its real part and imaginary part to calculate the amplitude. The process is as follows:

$$A(f) = \sqrt{\text{Re}[F(f)]^2 + \text{Im}[F(f)]^2} \quad (7)$$

where $\text{Re}[F(f)]$ is the real part of the fast fourier transform result $F(f)$, and $\text{Im}[F(f)]$ is the imaginary part of $F(f)$.

Step 3: Through peak analysis of the frequency spectrum, the frequency corresponding to the maximum peak is identified as the dominant frequency of the IMF component. Physically, a lower dominant frequency usually corresponds to long-term trend changes or base loads, while a higher frequency reflects rapid load fluctuations driven by short-term equipment operations or environmental changes.

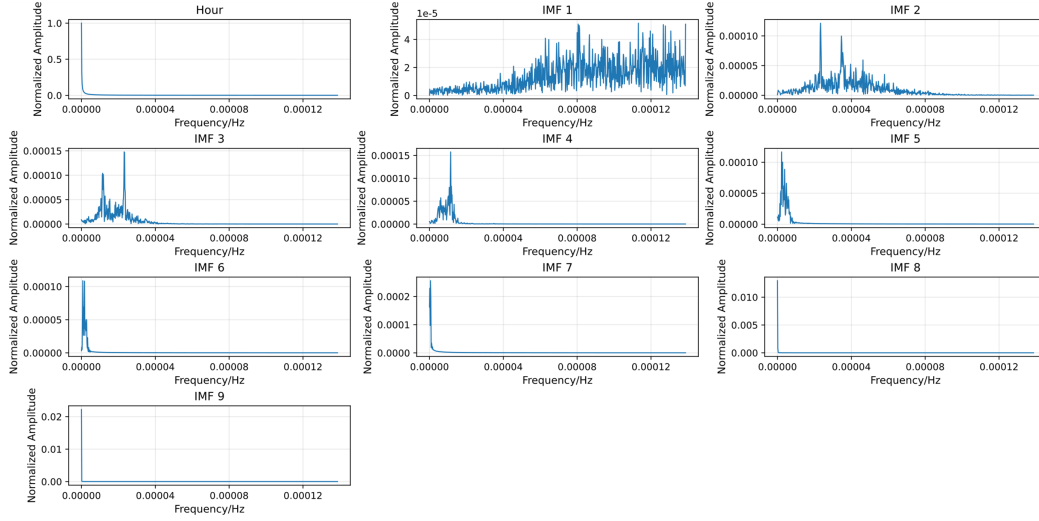


Figure 4: Spectrum analysis using FFT.

3.1.3 Complexity Analysis

Complexity is crucial in the process of signal analysis, especially for decomposed signals in electrical load forecasting. Thus, this paper uses sample entropy (SE) to calculate the complexity of each IMF component. The specific steps are as follows:

Step 1: For each IMF component $x_1, x_2, \dots, x_N$ decomposed by EEMD, take a subsequence of length m , defined as $x_i, x_{i+1}, \dots, x_{i+m-1}$, as a pattern.

Step 2: Define a distance metric, usually using the absolute difference to calculate the distance between subsequences. For each pair of subsequences (x_i, x_j) , if the distance between them is less than a given threshold r , they are considered similar. The distance d_{ij} between subsequences x_i and x_j is calculated as:

$$d_{ij} = \max_{0 \leq k \leq m-1} |x_{i+k} - x_{j+k}| \quad (8)$$

Step 3: According to the definition of sample entropy, calculate the sample entropy $SE(m, r)$ of the pattern, which measures the complexity of the sequence under a given length m and threshold r . The specific expression is as follows:

$$SE(m, r) = -\ln \left(\frac{A_m(r)}{B_m(r)} \right) \quad (9)$$

where $A_m(r)$ represents the number of all similar patterns, and $B_m(r)$ represents the number of all possible patterns.

Step 4: Calculate the sample entropy of all IMF components to obtain the complexity of each IMF component. The larger value indicates higher complexity of the IMF component. In the context of building electricity load, higher complexity (higher SE) indicates that the component contains more unpredictable and chaotic random events (such as sudden human activities or irregular plug loads), whereas lower complexity represents highly regular and predictable power consumption behaviors.

3.1.4 IMF Grouping Based on K-Means Clustering

Based on the main frequency and complexity extracted in the previous steps, this paper adopts the K-means clustering algorithm to automatically classify the decomposed IMF components. Specifically, for

the k -th IMF component, its dominant frequency f_k and sample entropy SE_k are combined to form a two-dimensional feature vector $v_k = [f_k, SE_k]$. To eliminate the dimensional difference between frequency and entropy, Min-Max normalization is first applied to the feature vectors. The objective of K-means is to partition the Kfeature vectors into $C = 3$ distinct clusters (corresponding to weak, medium, and strong components) $S = \{S_1, S_2, S_3\}$ by minimizing the within-cluster sum of squares (WCSS). The mathematical formulation of the clustering objective is defined as:

$$J = \sum_{c=1}^3 \sum_{v_k \in S_c} \|v_k - \mu_c\|^2 \quad (10)$$

where S_c represents the c -th cluster, and μ_c is the centroid of cluster S_c , which is updated iteratively by calculating the mean of all vectors assigned to that cluster:

$$\mu_c = \frac{1}{|S_c|} \sum_{v_k \in S_c} v_k \quad (11)$$

Through iterative optimization until the centroids converge, the IMFs are successfully classified. The classification results are presented in Fig. 5. In particular, different types of components correspond to different fluctuation conditions of electrical loads. Weak components usually have lower main frequencies and complexity, corresponding to the base load of buildings. It includes electricity required for the constant lighting in public areas of apartments and the operation of access control systems. Their electricity consumption is relatively stable, which can be used to predict basic electricity demand. Medium components have moderate main frequencies and complexity, associated with loads that operate periodically with daily routines (such as morning and evening washing), such as centralized hot water supply, which can be used to predict periodic demand. Strong components have higher main frequencies and relatively higher complexity, reflecting random load changes in buildings, which can be used to predict sudden demand.

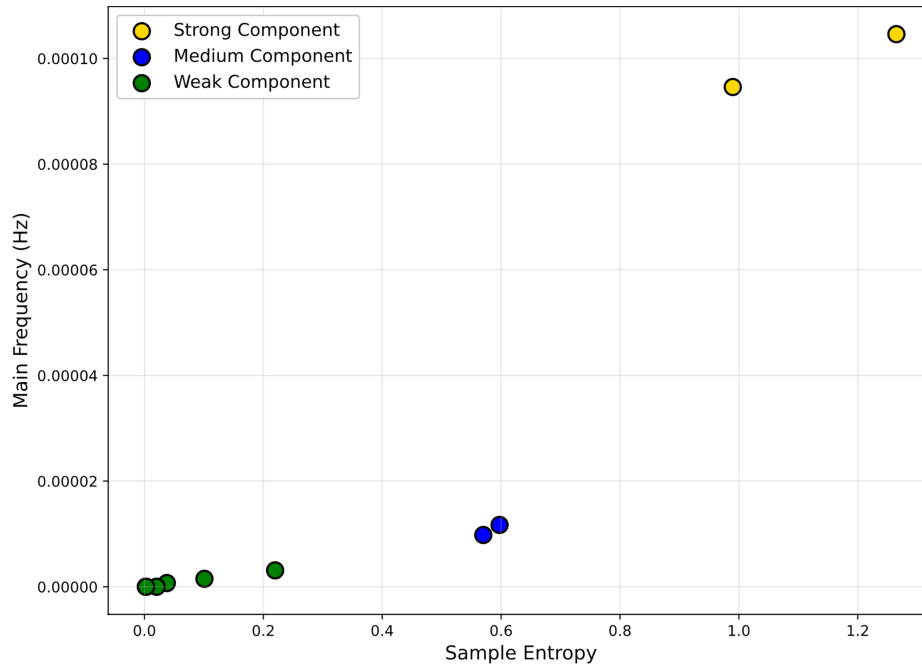


Figure 5: Results of K-means clustering.

3.2 Grouped Feature Attention Mechanism

BELF is influenced by historical load, external meteorological conditions, and calendar information, which increases the complexity of BELF. To address this issue, this paper develops a grouped feature attention mechanism (GFAM) that assigns appropriate weights to input features in order to identify key factors affecting BELF. The structure of the proposed grouped feature attention mechanism is presented in Fig. 6.

$$X_{B \times F_3 \times T}^{dw} = DWConv_{1 \times 3}(X_{B \times T \times F_3}) \quad (12)$$

$$X_{B \times 1 \times F_3}^{\max} = MaxPool(Permute(X_{B \times F_3 \times T}^{dw})) \quad (13)$$

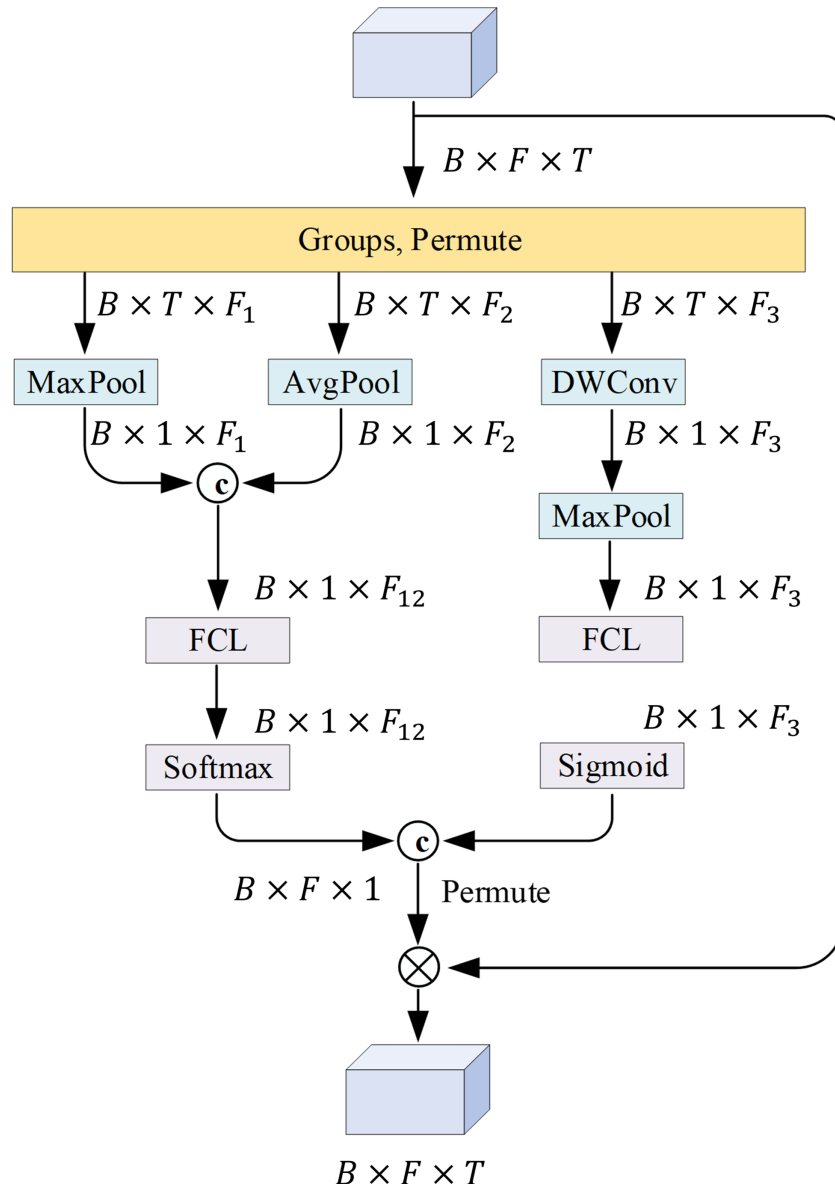


Figure 6: The structure of the grouped feature attention mechanism.

Given the input data $X_{B \times F \times T}$, where B is the batch size, F is the quantity of features, and T is the number of time steps. GFAM applies Group and Per-mute operations to divide the input into the load feature group

$X_{B \times T \times F_1}$, the meteorological feature group $X_{B \times T \times F_2}$, and the time feature group $X_{B \times T \times F_3}$, satisfying $F_1 + F_2 + F_3 = F$.

Then, GFAM applies max pooling and average pooling to the load feature group and the meteorological feature group, respectively. This helps extract peak information and trend information from the load sequence. The specific expressions are:

$$X_{B \times 1 \times F_1}^{\max} = \text{MaxPool}(X_{B \times T \times F_1}) \quad (14)$$

$$X_{B \times 1 \times F_2}^{\text{avg}} = \text{AvgPool}(X_{B \times T \times F_2}) \quad (15)$$

Next, the time feature group is first processed by depth wise convolution to capture local patterns, followed by max pooling for compression, as follows:

Then, GFAM concatenates the load group $X_{B \times 1 \times F_1}^{\max}$ and the meteorological group $X_{B \times 1 \times F_2}^{\text{avg}}$ to obtain $X_{B \times 1 \times F_{12}}$, which is adopted as the input of a fully connected layer (*FCL*) to capture the association between building load and meteorological conditions. Afterward, Softmax is applied to generate the attention weights $\alpha_{F_{12}}$. Meanwhile, GFAM feeds the time group $X_{B \times 1 \times F_3}^{\max}$ into another *FCL*, and applies Sigmoid to generate the attention weights α_{F_3} to extract the influence of time on BELF. The expressions are as follows:

$$X_{B \times 1 \times F_{12}} = \text{Concat}(X_{B \times 1 \times F_1}^{\max}, X_{B \times 1 \times F_2}^{\text{avg}}) \quad (16)$$

$$\alpha_{F_{12}} = \text{Softmax}(FCL(X_{B \times 1 \times F_{12}})) \quad (17)$$

$$\alpha_{F_3} = \text{Sigmoid}(FCL(X_{B \times 1 \times F_3}^{\max})) \quad (18)$$

Finally, $\alpha_{F_{12}}$ and α_{F_3} are concatenated to obtain the attention weights α_F covering all feature groups. The attention weights α_F are then multiplied element-wise with the input features $X_{B \times F \times T}$ to assign corresponding weights to the load, meteorological, and time groups, highlighting their contributions to BELF.

3.3 Multi-Scale Residual Depthwise Convolution Memory Module

This paper extracts a multi-scale residual depthwise convolution memory module (MRDCM) based on multi-scale residual depthwise convolution and gated recurrent unit to capture the long-term and short-term features and time patterns of building electrical load. The specific structure of MRDCM is depicted in Fig. 7.

3.3.1 Multi-Scale Residual Depthwise Convolution

CNN-based methods often extract general patterns of electrical load through single-scale convolution kernels. However, the above models cannot extract the change trend of building load under multiple scales, which limits the prediction ability of CNN-based methods. Therefore, this paper designs a multi-scale depthwise convolution module (MSDC) to enhance the spatial convolution ability, and its structure is depicted in Fig. 7.

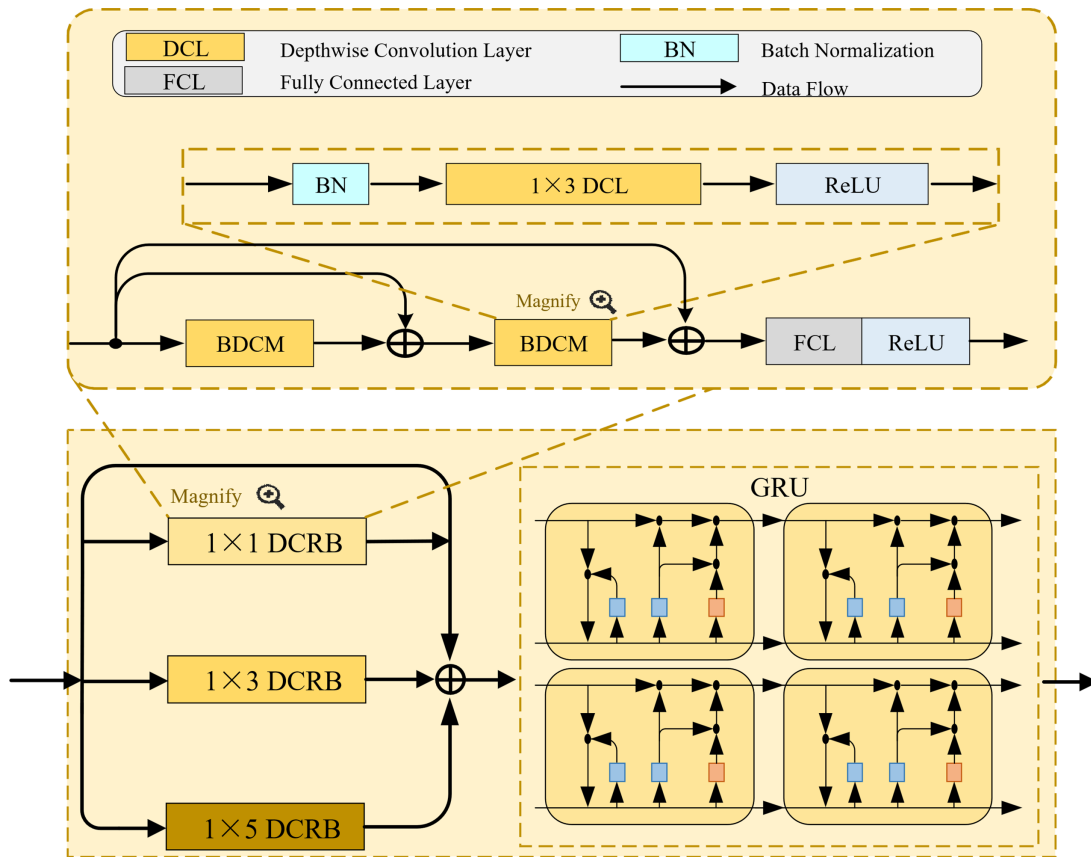


Figure 7: The structure of the multi-scale residual depthwise convolution memory module.

First, this paper develops a basic depthwise convolution block (BDCM). It comprises a depthwise convolution layer, batch normalization (BN), and ReLU activation function. Compared with conventional convolution, depthwise convolution performs filtering on each channel independently. This mechanism allows the model to capture channel-specific spatial information with much lower computational complexity. The computation process is described as follows:

$$Y_{B \times F \times T}^D = \text{ReLU} \left(\text{BN} \left(W_k * X_{B \times F \times T}^{\text{output}} + b_m \right) \right) \quad (19)$$

where $Y_{B \times F \times T}^D$ is the output of BDCM; w_k and b_m are the weight parameter and bias term of depthwise convolution, respectively; $*$ represents the standard depthwise convolution operation; $\text{BN}(\cdot)$ represents batch normalization operation, which serves to accelerate network convergence and prevent overfitting.

Then, by cascading and expanding BDCM with the residual convolution structure, a depthwise convolution layer residual block (DCRB) is developed and used as a sub-module of MSDC. Among them, residual connection can effectively alleviate the gradient disappearance problem and improve the model's feature representation capability. $\text{BDCM}(\cdot)$ denotes the calculation of the basic depthwise convolution block. The calculation process of DCRB is described as follows:

$$Y_{D1} = \text{BDCM} \left(X_{B \times F \times T}^{\text{output}} \right) \quad (20)$$

$$Y_{D2} = \text{BDCM} \left(Y_{D1} + X_{B \times F \times T}^{\text{output}} \right) \quad (21)$$

$$Y_D = Y_{D2} + X_{B \times F \times T}^{\text{output}} \quad (22)$$

Finally, this paper constructs the MSDC module by combining DCRBs of different scales. DCRB has convolution kernels of different sizes, $DCRB_{1 \times 1}$, $DCRB_{1 \times 3}$, and $DCRB_{1 \times 5}$ represent small-scale, medium-scale, and large-scale convolution kernels, respectively. $DCRB_{1 \times 1}$ and $DCRB_{1 \times 3}$ are used to extract short-term features of building load, and $DCRB_{1 \times 5}$ is used to learn long-term features of load. The calculation process of MRDCM is as follows:

$$Y_{o1} = DCRB_{1 \times 1} (X_{B \times F \times T}^{output}) \quad (23)$$

$$Y_{o2} = DCRB_{1 \times 3} (X_{B \times F \times T}^{output}) \quad (24)$$

$$Y_{o3} = DCRB_{1 \times 5} (X_{B \times F \times DT}^{output}) \quad (25)$$

$$Y_o = Y_{o1} + Y_{o2} + Y_{o3} + X_{B \times F \times T}^{output} \quad (26)$$

This multi-scale feature fusion aims to enable the model to have the perception ability of load changes in multiple time scales at the same time, thereby improving the prediction performance.

3.3.2 Gated Recurrent Unit Calculation Steps

The gated recurrent unit (GRU) is a variant of recurrent neural network (RNN), especially suitable for processing building thermal load data, and can capture the temporal dependency in the data. In this model, GRU takes the output of the multi-scale depthwise convolution module (MSDC) described in the previous section as input, so as to realize the modeling of multi-scale temporal dynamics in the load sequence. The calculation of GRU at each time step can be decomposed into the following four sequential steps:

Step 1: Update gate calculation. The update gate can control the fusion proportion of historical information and current information. This step is calculated as follows:

$$z_t = \sigma (W_z \cdot [h_{t-1}, x_t]) \quad (27)$$

where σ is the sigmoid activation function, and the output is between 0 and 1; W_z is the weight matrix of the update gate; h_{t-1} is the hidden state at the previous time step; x_t is the input at the current time step, that is, the multi-scale features extracted by the MSDC module.

Step 2: Reset gate calculation. The reset gate determines how much information in the previous hidden state h_{t-1} should be retained or ignored to better capture temporal features in combination with the current input x_t . The calculation formula is:

$$r_t = \sigma (W_r \cdot [h_{t-1}, x_t]) \quad (28)$$

where W_r is the weight matrix of the reset gate.

Step 3: Candidate hidden state calculation. The candidate hidden state $\tilde{h}_t$ integrates the current input and the historical information filtered by the reset gate, and is calculated in the following way:

$$\tilde{h}_t = \tanh (W \cdot [r_t * h_{t-1}, x_t]) \quad (29)$$

where W is the weight matrix, and $\tanh$ is the activation function, which is used to normalize the output to the interval $[-1, 1]$.

Step 4: Current hidden state update. The final hidden state h_t at the current time step is obtained by weighted fusion of the previous state h_{t-1} and the candidate state $\tilde{h}_t$ through the update gate:

$$h_t = z_t * h_{t-1} + (1 - z_t) * \tilde{h}_t \quad (30)$$

When z_t is close to 1, the model tends to retain more historical information; when it is close to 0, it focuses on current input information.

Through the above four sequential steps, GRU can effectively fuse the multi-scale spatial features provided by the MSDC module, realize the unified modeling of long-term and short-term temporal dependency in the building thermal load sequence, and then improve the prediction accuracy.

4 Experiment

This experiment uses two campus buildings to evaluate the prediction accuracy of MCG-GFAM-MRDCM. The experimental platform adopts Intel Core i9-13900HX, NVIDIA GeForce RTX 4070 Laptop GPU, and 32 GB dual-channel RAM. At the software level, deployment is completed through Python 3.7 and PyTorch framework.

4.1 Dataset Description and Experimental Setup

This paper uses the load data of university church and dormitory to verify the predictive performance of the proposed model. The dataset time range is from 00:00 on 1 January 2022 to 23:00 on 31 December 2024, with a time resolution of 1 h. The electricity load curve is shown in Fig. 8. Meteorological data is from the National Renewable Energy Laboratory. The training set, validation set, and test set are divided into 8:1:1.

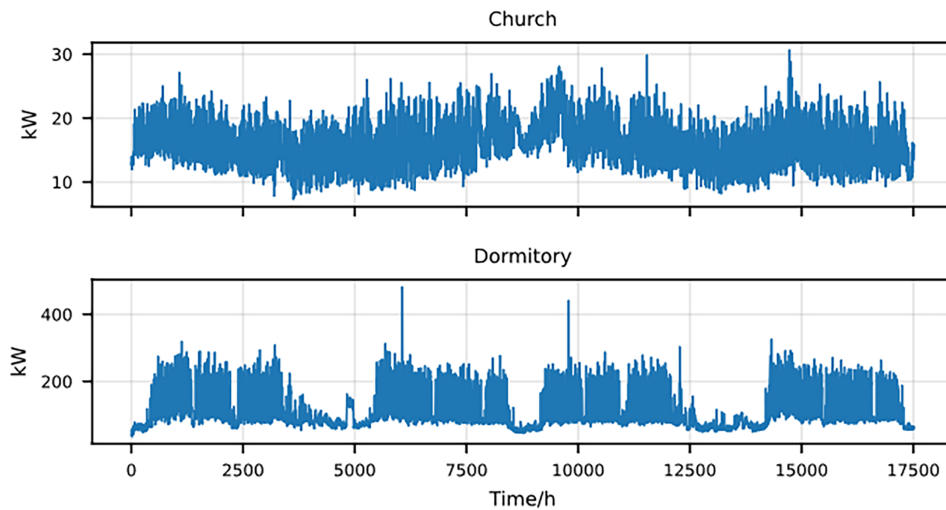


Figure 8: The electricity load curves of church and dormitory from 2022 to 2024.

To clarify the operational context in practical building energy management systems, this study adopts an offline training with online rolling forecasting paradigm. Let t denote the current time step. In this study, the length of the historical look-back window T is set to 24 h. At any given time t , the proposed MCG-GFAM-MRDCM model receives a multivariate input sequence $X_{t-23:t}$, which consists of historical building electricity load, meteorological data (humidity, temperature, pressure), and calendar information (time, holiday) over the past 24 h. The forecasting horizon is strictly defined as 1 h ahead, meaning the model

predicts the electricity load for time step $t + 1$. During the operational testing phase, an online rolling window mechanism is employed: as time moves forward, the sliding window shifts by 1 h, continuously discarding the oldest data point and incorporating the latest actual sensor observations to update the input sequence and generate the next prediction. This rolling mechanism ensures that the model rapidly adapts to dynamic and non-stationary operating conditions, making it highly suitable for integration with real-time energy management frameworks.

4.2 Model Hyperparameters

This paper adopts the grid search approach to optimize the hyperparameters of MCG-GFAM-MRDCM, as shown in Table 1. First, hyperparameters that affect the forecasting performance of the method are selected based on experience, and their value ranges are defined. Then, this part uses the grid search algorithm to find the optimal hyperparameters. Among them, (MCG) corresponds to the hyperparameters of the modal component grouping module, (GFAM) corresponds to the hyperparameters of the grouped feature attention mechanism module, and (MRDCM) corresponds to the hyperparameters of the multi-scale residual depthwise convolution memory module. In addition, the model selects Adam as the optimizer and takes the MAPE as the loss function. Moreover, the main hyperparameters of the comparison model are described in Table 2.

Table 1: Hyperparameters of the MCG-GFAM-MRDCM model.

Hyperparameters Search Space	Results
noise standard deviation (EEMD) = [0.05, 0.1, 0.15, 0.2]	0.1
Layers (GFAM) = [1, 2, 3, 4, 5]	3
kerne size (GFAM) = [1, 2, 3, 4, 5]	3
kerne size (MRDCM) = [1, 2, 3, 4, 5]	1, 3, 5
Layers (GRU) = [1, 2, 3, 4, 5]	3
Units (GRU) = [8, 16, 32, 64]	16, 8, 8
learning rate = [0.1, 0.01, 0.001, 0.0001]	0.01
batch size = [8, 16, 32, 64, 128]	16

Table 2: Hyperparameters of existing models.

Methods	Main Hyperparameters
CNN	Filter = 16, kernel size = 3, learning rate = 0.01, batch size = 16
GRU	Layers = 3, units = 8, 16, 32, learning rate = 0.001, batch size = 32
CNN-GRU	Filter = 16, kernel size = 5, layers = 3, units = 8, 16, 32 learning rate = 0.001, batch size = 16
CNN-GRU-AM	Filter = 8, kernel size = 3, layers = 2, units = 16, 32 learning rate = 0.001, batch size = 32

4.3 Performance Index

The experiments adopt the MAE, MAPE, and RMSE as evaluation metrics to validate the forecasting performance of the MCG-GFAM-MRDCM. Specifically, they are defined as follows:

$$MAE = \frac{1}{M} \sum_{s=1}^M |y_s - \hat{y}_s| \quad (31)$$

$$MAPE = \frac{1}{M} \sum_{s=1}^M \left| \frac{y_s - \hat{y}_s}{y_s} \right| \quad (32)$$

$$RMSE = \sqrt{\frac{1}{M} \sum_{s=1}^M (y_s - \hat{y}_s)^2} \quad (33)$$

where y_s represents the actual load at the s -th moment, $\hat{y}_s$ is the predicted load at the s -th moment, and M represents the amount of samples.

4.4 Ablation Study

4.4.1 Ablation Study of the MCG Module

To verify the effectiveness of the modal component grouping method (MCG), this experiment analyzes the role of two core steps: frequency domain analysis and complexity analysis in improving the rationality of modal grouping and prediction performance. For this reason, this part constructs two comparison models. MCG-FFT means removing frequency domain analysis from the MCG method and only relying on complexity analysis for grouping. MCG-SE means removing complexity analysis during the grouping process and only dividing based on frequency domain features. We perform predictions on church and dormitory datasets respectively, and use MAPE, MAE, and RMSE as evaluation indicators.

Fig. 9 shows the comparison of prediction errors of each model on the two datasets. The experimental results show that the MCG-GFAM-MRDCM model achieves the best performance on both datasets. Taking the church dataset as an example, compared with MCG-FFT, MCG-GFAM-MRDCM reduces MAPE by 0.73%, MAE by 0.15 kW, and RMSE by 0.08 kW. Compared with MCG-SE, MCG-GFAM-MRDCM reduces MAPE by 0.39%, MAE by 0.11 kW, and RMSE by 0.07 kW. A similar trend is observed on the dormitory dataset. Specifically, compared with MCG-FFT, MCG-GFAM-MRDCM reduces MAPE by 1.19%, MAE by 0.23 kW, and RMSE by 0.52 kW, and by 0.72%, 0.19, and 0.59 kW respectively compared with MCG-SE. The significant performance drop in both MCG-FFT and MCG-SE clearly justifies the physical necessity of our three-dimensional decomposition framework. Relying solely on complexity and time-domain (MCG-FFT) causes the model to miss the exact periodicity of user behaviors (e.g., daily operating schedules of HVAC systems), potentially misclassifying a highly regular but frequent fluctuation as random noise. Conversely, relying solely on the frequency and time-domain (MCG-SE) makes it difficult for the model to distinguish between a high-frequency regular signal and high-frequency chaotic noise (like irregular random plug loads), since both share similar frequency ranges but possess vastly different unpredictability. Therefore, the superior forecasting performance of MCG-GFAM-MRDCM stems from the synergistic fact that: EEMD separates the scales in the time domain, FFT identifies the periodic patterns in the frequency domain, and SE filters the randomness based on complexity. This three-dimensional strategy allows the subsequent K-means clustering to accurately map the mathematical IMFs into physically meaningful load components (stable base loads, periodic loads, and sudden random loads). By perfectly aligning the data characteristics with the physical realities of building energy consumption, it provides highly targeted inputs for the deep learning modules, thereby fundamentally improving the overall prediction accuracy.

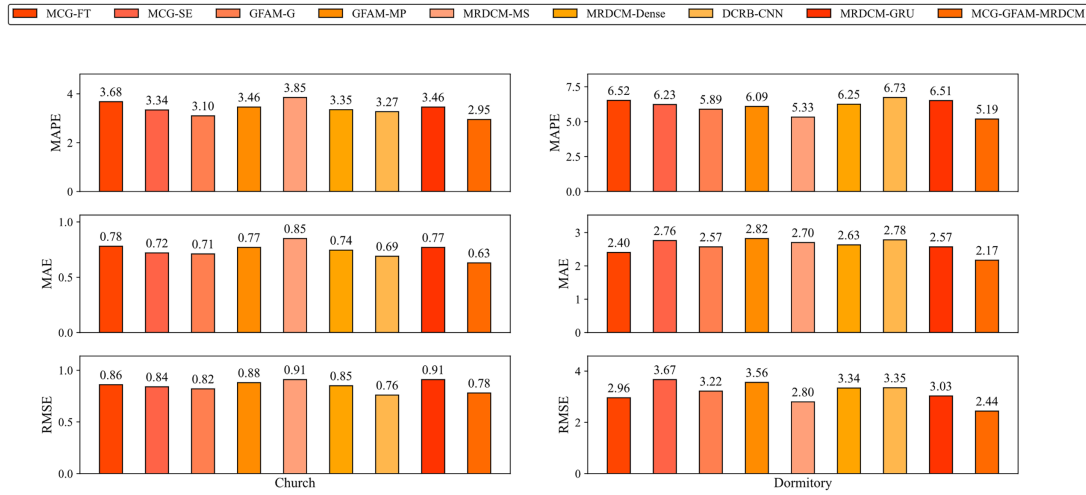


Figure 9: Results of the ablation study evaluating the contribution of different modules.

4.4.2 Ablation Study of the GFAM Module

To verify the role of feature grouping and maximum pooling operations in the grouped feature attention mechanism (GFAM) in key input features, this experiment constructs two comparison models including GFAM-G (removing feature grouping operations) and GFAM-MP (removing maximum pooling operations). Subsequently, the performance of the comparison models was evaluated on two datasets, and the experimental results are shown in Fig. 9. It can be seen that MCG-GFAM-MRDCM with the complete GFAM module shows better prediction performance on both datasets. On the church dataset, compared with GFAM-G, the MAPE of MCG-GFAM-MRDCM is reduced by 0.15%, MAE by 0.08 kW, and RMSE by 0.08 kW. The results show that the feature grouping operation can cluster input features with similar semantics (such as different types of loads, multiple meteorological parameters) to reduce interference between irrelevant features, enabling MCG-GFAM-MRDCM to more accurately extract associated information between similar features. Theoretically and statistically, this superiority of the grouping strategy over standard flat attention schemes stems from two aspects. First, it prevents weight suppression among heterogeneous data. The input variables (historical loads, meteorological data, and calendar information) possess completely different numerical scales and statistical variances. If all features are directly fed into a global attention mechanism, those with large numerical variances (such as highly fluctuating load components) mathematically dominate the Softmax probability distribution, effectively suppressing the attention weights of discrete or low-variance features (such as holiday indicators). GFAM resolves this by evaluating importance within independent groups. Second, grouping reduces meaningless statistical correlations. In a standard un-grouped attention scheme, the model often incorrectly links a random spike in historical load to an unrelated weather variable simply because they co-occur in the training set. By explicitly establishing boundaries between the load, meteorological, and time domains, GFAM filters out irrelevant noise, reduces the complexity of feature interactions, and makes the forecasting model more stable and interpretable. Compared with GFAM-MP, the MAPE of MCG-GFAM-MRDCM is further reduced by 0.12%, MAE by 0.11 kW, and RMSE by 0.10 kW. It indicates that the maximum pooling operation can effectively increase the weight proportion of key features in the group, weaken the impact of redundant information, and make feature screening more targeted. On the dormitory dataset, compared with GFAM-G, MCG-GFAM-MRDCM reduces MAPE by 0.56%, MAE by 0.30 kW, and RMSE by 0.78 kW. Compared with GFAM-MP,

the three indicators are reduced by 0.16%, 0.13, and 0.9 kW, respectively. It verifies that the effectiveness of GFAM has good stability under different datasets.

This paper further analyzes the attention changes of different types of features by visualizing the dynamic evolution of attention weights during the training process of the MCG-GFAM-MRDCM model, as shown in Fig. 10. The features include different types of load components, meteorological features, and time features. Experiments were conducted on the church and dormitory datasets respectively to verify the stability of GFAM. In the early stage of training, the attention distribution of MCG-GFAM-MRDCM is relatively scattered on both datasets. Although load components occupied a certain advantage, the overall weights fluctuated significantly. Time information also played a certain role. However, the weights of meteorological features and holiday features are relatively low, and the model has not yet formed a clear feature preference. As the number of iterations increases, MCG-GFAM-MRDCM gradually captures the connections between features. On the church dataset, the weights of medium component and weak component keep rising, and the importance of Humidity increases significantly during the fluctuation stage. On the dormitory dataset, the weight of strong component increases particularly obviously. The Holiday feature shows a higher attention proportion, while the weight of temperature remains at an extremely low level.

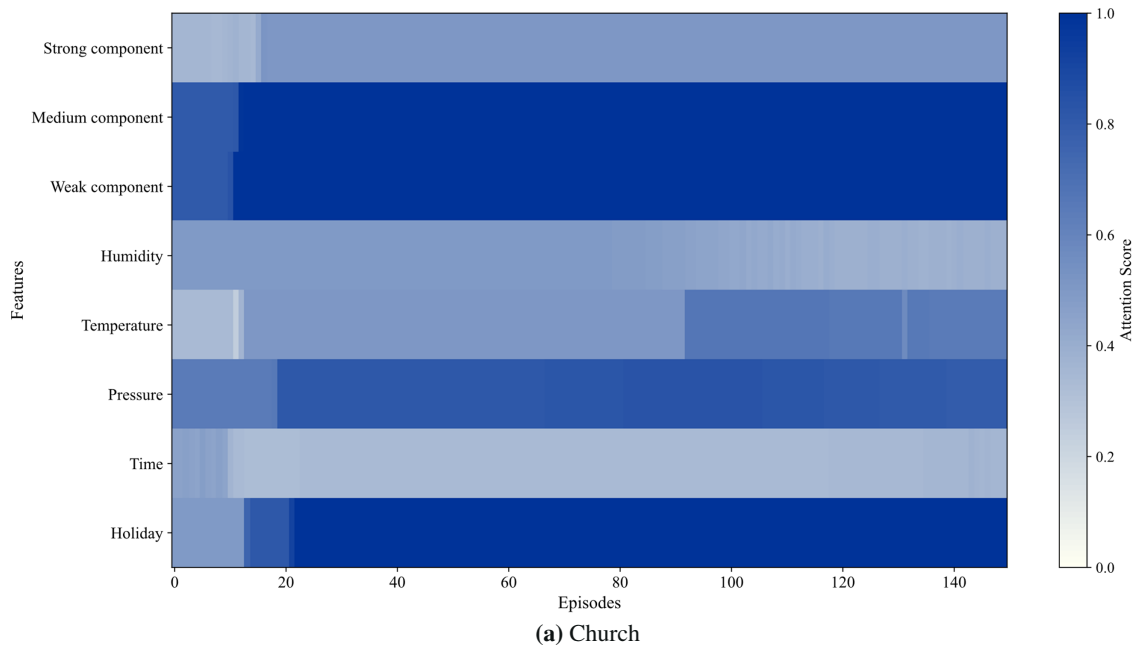


Figure 10: (Continued)

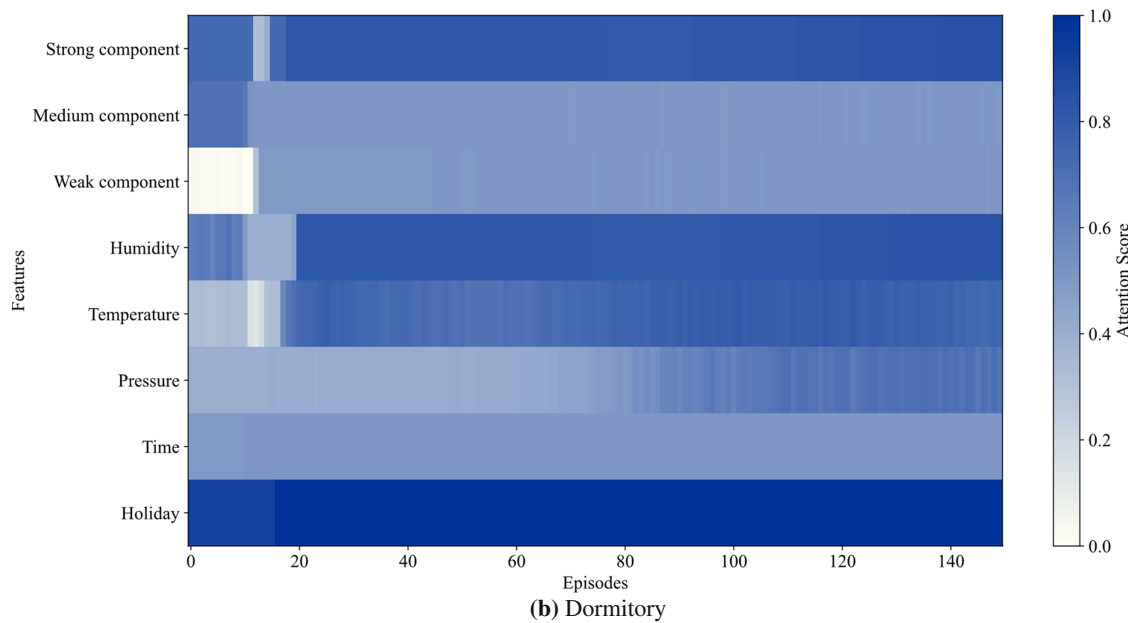


Figure 10: The iterative convergence process of the attention weight.

When the model converges, it exhibits a stable weight ranking on both datasets, as shown in Fig. 11. Specifically, on the church dataset, strong component (average 0.9631), humidity (average 0.9595), and weak component (average 0.9423) occupy a dominant position. The average attention scores of the three all exceed 0.94. The weights of temperature (average 0.4924) and time (average 0.4606) are relatively low, below 0.5. On the dormitory dataset, time (average 0.9475) and strong component (average 0.8834) are the two most important features. Their average attention scores approach or exceed 0.88. The weight of temperature is extremely low, with an average of only 0.0204 and almost no impact. The weights of weak component (average 0.5233) and pressure (average 0.4763) are also at low level. This result indicates that there are significant differences in the key features affecting load prediction in different scenarios. Meteorological features, especially humidity, and load components have a more prominent impact in the church scenario. Time information and strong component play a more critical role in the dormitory scenario. In conclusion, MCG-GFAM-MRDCM shows a consistent attention change trend on both datasets and can dynamically identify the importance of different features in different time periods. The above results not only verify the rationality of GFAM but also prove that MCG-GFAM-MRDCM has good generalization ability and interpretability.

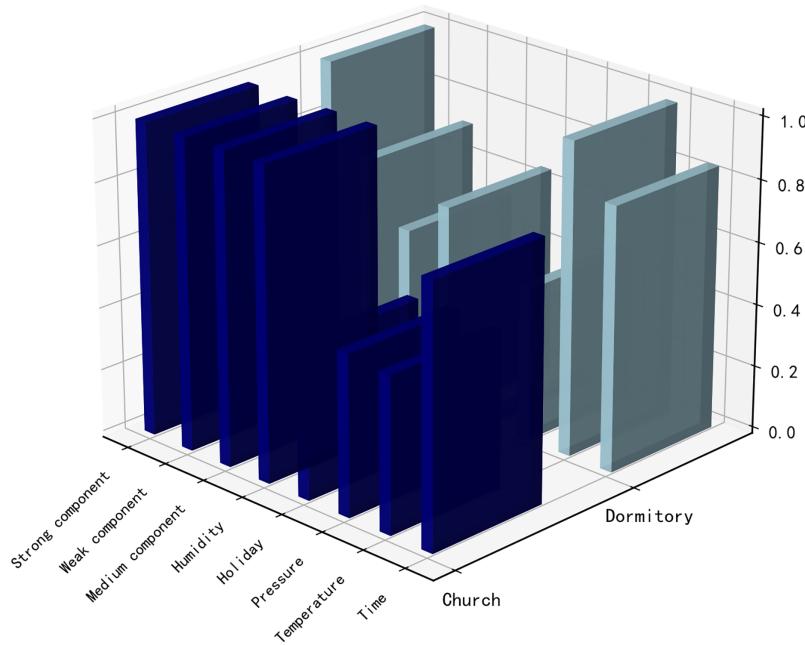


Figure 11: The weights of the grouped feature attention mechanism.

4.4.3 Ablation Study of the MRDCM Module

To verify the effectiveness of the multi-scale residual depthwise convolution memory module (MRDCM), this paper designs multiple comparison models to evaluate the role of each module in extracting long-term and short-term load features and time patterns. The comparison models include MRDCM-MS, MRDCM-Dense, DCRB-CNN, and MRDCM-GRU. MRDCM-MS means simplifying the multi-scale DCRB to a single-scale $\text{DCRB}_{1 \times 3}$. MRDCM-MS only uses 1×3 convolution kernels to extract load features. MRDCM-Dense replaces the residual connections in MRDCM with dense residual connections. DCRB-CNN replaces the depthwise convolution in DCRB with standard fixed convolution. MRDCM-GRU means removing the GRU module in MRDCM. Fig. 9 shows the prediction results of MCG-GFAM-MRDCM and the comparison models. It can be seen that compared with MRDCM-MS, on the church dataset, the MAPE of MRDCM is reduced by about 0.32%, RMSE by 0.08 kW, and MAE by 0.15 kW. On the dormitory dataset, MAPE is reduced by 0.04%, RMSE by 0.78 kW, and MAE by 0.40 kW. It indicates that multi-scale convolution can more comprehensively extract load features at different time scales. However, single-scale convolution has shortcomings in the richness of feature extraction, which limits the BELF accuracy. Compared with MRDCM-Dense, MRDCM reduces MAPE by 0.35%, RMSE by 0.04 kW, and MAE by 0.05 kW on the church dataset. On the dormitory dataset, MAPE is reduced by 0.64%, RMSE by 0.09 kW, and MAE by 0.05 kW. It indicates that residual connections have more advantages in ensuring effective information transmission and alleviating the gradient disappearance problem. Moreover, although dense residual connections can enhance feature reuse, they also bring additional redundant information. Compared with DCRB-CNN, MRDCM reduces RMSE by 0.04 kW, and MAE by 0.05 kW on the church dataset. On the dormitory dataset, MAPE is reduced by 0.32%, RMSE by 0.11 kW, and MAE by 0.05 kW. It shows that depthwise convolution can adaptively adjust the receptive field and better adapt to the irregular distribution of load data, while standard fixed convolution has poor flexibility in dealing with complex load forms. Compared with MRDCM-GRU, MRDCM reduces MAPE by 0.11%, RMSE by 0.03 kW, and MAE by 0.04 kW on the church dataset. On the

dormitory dataset, MAPE is reduced by 0.27%, RMSE by 0.07 kW, and MAE by 0.15 kW. It indicates that the GRU module plays a key role in capturing the long-term dependencies of load time series, and the lack of this module will limit the model's ability to learn long-term load patterns.

To intuitively present the performance advantages of the MCG-GFAM-MRDCM model, this paper compares and displays the 48-h load prediction results of MCG-GFAM-MRDCM with MRDCM-MS, MRDCM-Dense, DCRB-CNN, and MRDCM-GRU on two datasets, as shown in Fig. 12. Among them, the black line represents the actual electricity load, and the yellow line is the predicted value of MCG-GFAM-MRDCM. It can be seen that in the stable load periods of the two datasets, such as 1:00–7:00 on church dataset and 0:00–8:00 on dormitory dataset, all models can obtain accurate prediction results. However, when the load enters a period of large fluctuations, such as 10:00–14:00 on church dataset and 15:00–20:00 on dormitory dataset, the prediction performance of different models shows obvious differences. MRDCM-MS relies only on 1×3 convolution kernels to extract features, so its adaptability to fluctuating loads is insufficient, and the predicted values deviate greatly from the actual values. DCRB-CNN uses fixed convolution instead of depthwise convolution, making it difficult to capture dynamic features in load fluctuations. Although the dense residual connections of MRDCM-Dense enhance feature transmission, they are prone to redundancy when processing sharply changing load information. MRDCM-GRU lacks the ability of the GRU module to model time series dependencies, so it cannot accurately track the load fluctuation trend. However, MCG-GFAM-MRDCM can better fit the actual load through multi-scale feature extraction and effective module collaboration. This verifies the superiority of MCG-GFAM-MRDCM in complex load scenarios.

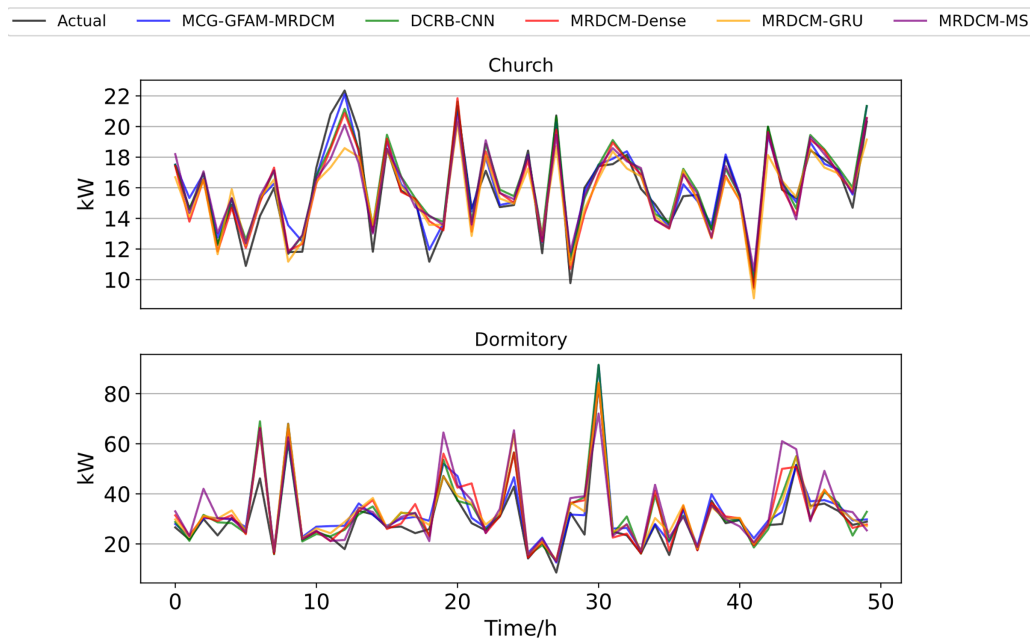


Figure 12: Electricity load forecasting curves with different models.

Beyond the quantitative error reductions, analyzing these ablation results reveals the critical physical role of the MRDCM module in handling diverse building energy behaviors. The electricity load of the church dataset is typically characterized by distinct, scheduled peak events interspersed with extended periods of stable base loads. The multi-scale architecture (validated by the performance drop in MRDCM-MS) perfectly adapts to this by utilizing larger convolution kernels to capture the stable baseline and smaller kernels to sharply pinpoint abrupt, event-driven spikes. In contrast, the dormitory dataset exhibits

highly stochastic and frequent load fluctuations driven by irregular student behaviors, such as the random usage of personal appliances. In this scenario, standard convolutions (as seen in DCRB-CNN) suffer from feature redundancy and struggle with irregular noise. Our adoption of depthwise convolution resolves this by processing channels independently, which significantly enhances the model's robustness against complex, high-frequency behavioral noise without overfitting. Finally, the integration of the GRU module is indispensable; it allows the model to continuously track the temporal momentum of these extracted spatial features, explaining why the complete MRDCM framework is essential for maintaining high prediction stability across entirely distinct operational building scenarios.

4.4.4 Comparison of Computational Efficiency

The training time of the BELF model directly affects its iteration efficiency in building energy management systems. Excessively long training time will prolong the model optimization cycle, making it difficult to quickly adapt to the dynamic changes in building electricity load. Therefore, this section analyzes the impact of each internal module of the MCG-GFAM-MRDCM model on the training time to evaluate the computational efficiency of the model, as shown in Table 3. Experiments were conducted on the church and dormitory datasets to compare the training times of the MCG-GFAM-MRDCM model and its various ablated versions. The training time of MCG-GFAM-MRDCM is 63.5 s on the church dataset and 66.2 s on the dormitory dataset. After removing the non-modal component grouping, GFAM-MRDCM eliminates feature processing steps such as EEMD decomposition, Fourier frequency domain analysis, and sample entropy complexity calculation. As a result, its training time is reduced to 48.2 and 50.5 s on the two datasets respectively, a decrease of 15.3–15.7 s compared with MCG-GFAM-MRDCM. MCG-MRDCM does not require feature grouping pooling and dynamic weight learning, so its training time drops to 56.8 and 59.1 s on the two datasets respectively, which is 6.7–7.1 s less than that of MCG-GFAM-MRDCM. The training time of MRDCM-MS is shortened to 59.2 and 61.7 s on the two datasets respectively, a reduction of 4.3–4.5 s compared with MCG-GFAM-MRDCM. This is because MRDCM-MS reduces the parameter computation of 1×1 and 1×5 convolution kernels. Due to the reduction of parameter iteration for temporal dependency modeling, the training time of MRDCM-GRU decreases to 55.1 and 57.5 s on the two datasets respectively, which is 8.4–8.7 s less than that of MCG-GFAM-MRDCM. In terms of the impact degree of each module on the training time, the MCG module contributes the most, followed by the GRU unit of MRDCM, while the GFAM module and the multi-scale characteristic of MRDCM have relatively smaller impacts. Although each module increases a certain amount of training overhead, they provide key support for the model to capture multi-dimensional load features and improve prediction accuracy through synergy. This design ultimately achieves a balance between training efficiency and prediction performance, which can meet the practical application requirements of building electricity load prediction.

Table 3: Comparison of model training time.

Methods	Training Time (s)	
	Church	Dormitory
MCG-GFAM-MRDCM	63.5	66.2
GFAM-MRDCM	48.2	50.5
MCG-MRDCM	56.8	59.1
MRDCM-MS	59.2	61.7
MRDCM-GRU	55.1	57.5

Based on the aforementioned ablation studies and training time evaluations, a further discussion on model capacity, overfitting risk, and parameter efficiency is necessary to justify the architectural complexity of the proposed MCG-GFAM-MRDCM relative to simpler temporal models. First, regarding model capacity, building electricity loads exhibit highly non-stationary characteristics with both short-term spatial spikes and long-term temporal trends. Simpler baseline models (such as pure CNN or GRU in [Section 4.5](#)) lack the representational capacity to model these dual dynamics simultaneously. The proposed hybrid architecture integrates multi-scale convolutions to capture local spatial features and a GRU to track temporal dependencies, significantly expanding the model's capacity to fit complex load patterns. Second, regarding overfitting risk, while complex models are typically prone to overfitting, our architecture introduces implicit structural regularizations to mitigate this issue. Specifically, the adoption of Depthwise Convolution in the MRDCM module drastically reduces the number of trainable parameters compared to standard convolutions, preventing the network from memorizing noise. Furthermore, the grouping strategy in the GFAM limits spurious statistical correlations between heterogeneous features. The consistently superior performance on the unseen test sets (as shown in [Fig. 13](#)) confirms that the model generalizes well without overfitting. Finally, regarding parameter efficiency, the trade-off between computational cost and accuracy is highly favorable. As shown in [Table 3](#), the inclusion of the GRU module (comparing MRDCM-GRU to the full model) only increases the training time by approximately 8.4 to 8.7 s. However, this marginal increase in computational overhead yields substantial reductions in prediction errors (e.g., a 0.27% reduction in MAPE on the Church dataset, as discussed in [Section 4.4.3](#)). This demonstrates that the carefully designed modules in MCG-GFAM-MRDCM achieve high parameter efficiency, making the added architectural complexity well justified for accurate building load forecasting tasks.

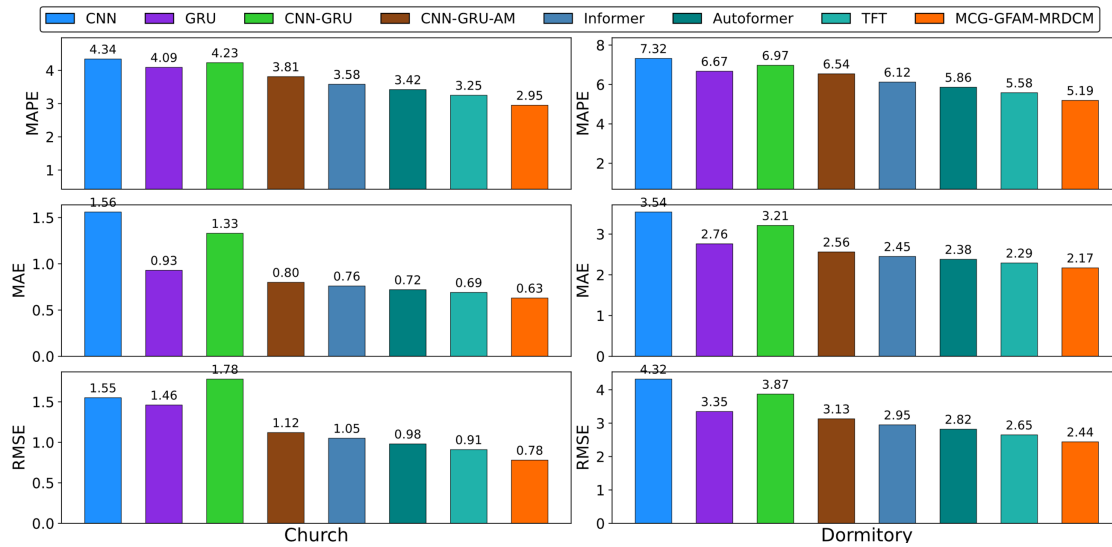


Figure 13: Comparison of the forecasted results of MCG-GFAM-MRDCM and existing methods.

4.4.5 Sensitivity Analysis of MCG Parameters

To verify the robustness of the Modal Component Grouping (MCG) method, this section conducts a sensitivity analysis on two core parameters: the number of clusters (K) in the K-means algorithm and the noise standard deviation (δ) in EEMD. The evaluations are performed on both the church and dormitory datasets, utilizing MAPE, RMSE, and MAE as metrics.

This part evaluates the model's forecasting performance by varying the number of grouped components $K \in \{2, 3, 4, 5\}$, as shown in Table 4. The results indicate that the optimal prediction performance is achieved at $K = 3$. When $K = 2$, the model suffers from under-segmentation, failing to effectively distinguish periodic loads from random sudden loads, which leads to an increase in MAPE by approximately 0.41% on the Church dataset. Conversely, when $K \geq 4$, over-segmentation occurs. The fragmented IMF components introduce redundant features that interfere with the grouped feature attention mechanism (GFAM), causing performance degradation. The mathematical optimum at $K = 3$ perfectly aligns with the physical semantics of building loads defined in this study: base load (weak), periodic demand (medium), and sudden demand (strong).

Table 4: Comparison of model training time.

Dataset	Number of Clusters (K)	MAPE (%)	RMSE (kW)	MAE (kW)
Church	K = 2	3.36	0.85	0.71
	K = 3	2.95	0.78	0.63
	K = 4	3.12	0.81	0.67
	K = 5	3.28	0.83	0.69
Dormitory	K = 2	5.82	2.65	2.38
	K = 3	5.19	2.44	2.17
	K = 4	5.46	2.51	2.25
	K = 5	5.61	2.58	2.31

The amplitude of the added white noise δ controls the mitigation of mode mixing during the EEMD process. This section analyzes the sensitivity of δ within the search space of $[0.05, 0.1, 0.15, 0.2]$, with results presented in Table 5. When $\delta = 0.05$, the noise is insufficient to suppress mode mixing entirely, slightly increasing the prediction error. Conversely, when $\delta \geq 0.15$, excessive white noise generates redundant high-frequency intrinsic mode functions (IMFs), which increases noise interference and limits the accuracy of BELF. The model achieves peak performance at $\delta = 0.1$. Notably, despite the variations in δ , the fluctuations in MAPE remain relatively small, indicating that the proposed framework possesses robust generalization capabilities against decomposition parameter variations.

Table 5: Forecasting performance under different EEMD noise standard deviations (δ).

Dataset	Noise Level (δ)	MAPE (%)	RMSE (kW)	MAE (kW)
Church	$\delta = 0.05$	3.15	0.82	0.68
	$\delta = 0.10$	2.95	0.78	0.63
	$\delta = 0.15$	3.09	0.80	0.66
	$\delta = 0.20$	3.24	0.84	0.70
Dormitory	$\delta = 0.05$	5.58	2.55	2.29
	$\delta = 0.10$	5.19	2.44	2.17
	$\delta = 0.15$	5.37	2.49	2.22
	$\delta = 0.20$	5.65	2.59	2.33

4.4.6 Statistical Significance and Uncertainty Analysis

To rigorously validate the forecasting superiority and stability of the proposed MCG-GFAM-MRDCM, this section conducts statistical significance testing and uncertainty analysis. First, the diebold-mariano (DM) test is employed to evaluate whether the performance improvements over the baseline models are statistically significant. The null hypothesis (H_0) assumes no significant difference in predictive accuracy. As presented in Table 6, all DM statistics are significantly large positive values, and all corresponding p -values are strictly below the 0.001 threshold. Consequently, H_0 is strongly rejected at the 1% significance level for all comparisons, providing solid statistical evidence that the model's superiority is genuine rather than coincidental.

Table 6: Diebold-Mariano (DM) test results comparing MCG-GFAM-MRDCM with other methods.

Dataset	Models	DM Statistic	p -Value	Significance Level
Church	DCRB-CNN	17.593	<0.001	1%
	MRDCM-MS	17.974	<0.001	1%
	MRDCM-Dense	11.464	<0.001	1%
	MRDCM-GRU	16.686	<0.001	1%
Dormitory	DCRB-CNN	8.609	<0.001	1%
	MRDCM-MS	15.295	<0.001	1%
	MRDCM-Dense	13.089	<0.001	1%
	MRDCM-GRU	8.693	<0.001	1%

To quantify the forecasting uncertainty, this study calculates the 95% confidence interval (CI) of the absolute errors (AE) for each model, as summarized in Table 7. A narrower 95% CI indicates lower uncertainty and higher prediction reliability. The results demonstrate that the proposed MCG-GFAM-MRDCM not only achieves the lowest mean absolute error but also maintains the narrowest 95% CI bounds on both datasets. For instance, the CI width of the proposed model is only 0.03 kW ([0.51, 0.54]) on the church dataset. This tighter error distribution implies that the model effectively suppresses extreme outlier predictions and exhibits the highest stability for practical building energy management.

Table 7: Uncertainty analysis based on the 95% Confidence Interval (CI).

Dataset	Model	95% CI
Church	DCRB-CNN	[0.72, 0.78]
	MRDCM-MS	[0.75, 0.81]
	MRDCM-Dense	[0.65, 0.71]
	MRDCM-GRU	[0.75, 0.81]
	MCG-GFAM-MRDCM	[0.51, 0.54]
Dormitory	DCRB-CNN	[3.89, 4.29]
	MRDCM-MS	[5.10, 5.68]
	MRDCM-Dense	[4.60, 5.09]
	MRDCM-GRU	[3.77, 4.11]
	MCG-GFAM-MRDCM	[3.15, 3.46]

4.5 Comparison with Existing Models

To evaluate the performance of the MCG-GFAM-MRDCM model in the building electricity load forecasting (BELF) task, this paper compares it with seven benchmark models: CNN, GRU, CNN-GRU, CNN-GRU-AM, and three state-of-the-art (SOTA) Transformer-based models (Informer, Autoformer, and Temporal Fusion Transformer (TFT)) on two real building load datasets, as shown in Fig. 13. All models use the same input settings, and the optimal hyperparameters are determined through grid search to ensure the fairness of the comparison. The experimental results show that MCG-GFAM-MRDCM achieves the best performance in terms of MAPE, MAE, and RMSE indicators on both datasets, significantly outperforming other models. Specifically, in the church dataset, MCG-GFAM-MRDCM has the lowest MAPE, MAE, and RMSE. Compared with the best-performing SOTA baseline model, TFT, the MAPE is reduced by 0.30%, MAE by 0.06 kW, and RMSE by 0.13 kW. Compared with CNN-GRU-AM, the MAPE is reduced by 0.86%, MAE by 0.17 kW, and RMSE by 0.34 kW. This indicates that MCG-GFAM-MRDCM has more advantages in capturing load time series features and changing trends. In the dormitory dataset, MCG-GFAM-MRDCM also maintains the best performance. Compared with TFT, the average MAPE is reduced by 0.39%, the average MAE by 0.12 kW, and the average RMSE by 0.21 kW. This verifies the stability and generalization ability of MCG-GFAM-MRDCM in different building scenarios. It can be seen from Fig. 13 that CNN has the highest error. This indicates that relying solely on convolution to extract local patterns while ignoring time dependencies will lead to insufficient characterization of dynamic load changes.

Although GRU can model the time dependence of loads, it cannot extract spatial information of loads. Therefore, its performance is better than CNN, but the improvement is limited. Compared with CNN and GRU, CNN-GRU has lower prediction errors. This is because CNN-GRU combines convolution and recurrent units, which can capture both spatial and temporal features of loads. CNN-GRU-AM introduces an attention mechanism on the basis of CNN-GRU, which can highlight the impact of key moments on BELF. Notably, the Transformer-based models (Informer, Autoformer, and TFT) exhibit strong competitiveness, significantly outperforming traditional deep learning models. TFT, in particular, achieves the second-best prediction performance due to its specialized variable selection network. However, standard Transformer architectures process raw sequence data directly, which often causes their self-attention mechanisms to be distracted by high-frequency noise and mode mixing inherent in building loads. Furthermore, while they excel at modeling long-range dependencies, they often struggle to capture highly localized spatial sudden spikes. In contrast, MCG-GFAM-MRDCM achieves the best performance. This is because it achieves a deeper integration in feature modeling strategies. First, the modal component grouping method finely decomposes the original consumption signal through frequency-domain analysis and complexity analysis, effectively distinguishing patterns of different scales and complexities, and reducing noise interference. Secondly, the grouped feature attention mechanism performs grouping processing on features and uses maximum pooling to strengthen key information, achieving more accurate attention allocation. Finally, the MRDCM module explicitly fuses multi-receptive-field local features with temporal dependencies, adapting to dynamic changes in loads more effectively than flat Transformer models. Therefore, MCG-GFAM-MRDCM significantly outperforms the SOTA comparison models in performance. In conclusion, MCG-GFAM-MRDCM can achieve more efficient load prediction in complex building scenarios.

Furthermore, to comprehensively evaluate the models' suitability for online operational environments, we compared their testing times (inference times) on the testing sets, as presented in Table 8. Testing time is a critical metric for real-time building energy management systems. As shown, while lightweight models like CNN and GRU exhibit the lowest testing times (ranging from 0.358 to 0.757 s), their prediction accuracies are significantly inferior. Notably, the proposed MCG-GFAM-MRDCM requires only 0.843 and 0.962 s on the Church and Dormitory datasets, respectively. More importantly, it demonstrates faster inference speeds

compared to advanced SOTA baselines such as Autoformer 0.884/0.958 s and TFT 0.956/1.028 s, successfully avoiding the heavy computational burden typical of global self-attention mechanisms.

Table 8: Comparison of model testing time.

Methods	Testing Time (s)	
	Church	Dormitory
CNN	0.452	0.358
GRU	0.658	0.757
CNN-GRU	0.772	0.628
CNN-GRU-AM	0.585	0.658
Informer	0.805	0.852
Autoformer	0.884	0.958
TFT	0.956	1.028
MCG-GFAM-MRDCM	0.843	0.962

Despite achieving a superior balance between forecasting accuracy and testing efficiency, this comparative analysis also highlights certain limitations of the proposed framework. First, compared to end-to-end basic models (e.g., CNN or GRU), our model requires a separate offline preprocessing phase. The Modal Component Grouping (MCG) mechanism, relying on EEMD and iterative K-means clustering, inevitably introduces algorithmic complexity before the deep learning forecasting begins. Second, the robust performance of the multi-scale depthwise convolution and GRU modules is highly dependent on the continuity and quality of historical multi-variate data. In practical scenarios involving frequent sensor communication failures or large segments of missing data, the model may be more susceptible to error propagation than simpler statistical methods. Finally, while meteorological and calendar features are integrated, the current framework does not incorporate stochastic human behavioral inputs—such as real-time building occupancy or subjective thermal adjustments—which remain unpredictable factors causing sudden load spikes in all compared models.

5 Conclusion

This paper proposes a novel deep learning method, denoted as MCG-GFAM-MRDCM, for capturing the dynamic changes in building loads and the influence of external factors on the load, thereby enhancing the model's generalization ability. The research conclusions are outlined as follows:

- (1) The proposed modal component grouping (MCG) method can decompose the input load signal into weak, medium, and strong IMF components to extract frequency-domain information and complex fluctuation characteristics of the load.
- (2) The proposed multi-scale depthwise convolutional memory (MRDCM) module can extract short-term features under building load fluctuations and also capture long-term trends during stable load states. The introduction of GRU further enhances the model's ability to capture temporal patterns of the load.
- (3) The proposed grouped feature attention mechanism (GFAM) can automatically assign weights to exogenous variables to identify the dominant factors affecting building load forecasting, thereby improving feature extraction efficiency.

- (4) Compared with existing models, the proposed MCG-GFAM-MRDCM can output the most accurate prediction results. In terms of MAPE, MAE, and RMSE performance metrics, it achieved reductions of 0.9%, 1.08%, and 0.4%, respectively.

In future research, we plan to further extend the work in the following directions:

- (1) The current model primarily considers the impact of meteorological and historical load factors on building electricity load. Future work could analyze building energy use behavior characteristics (such as occupancy patterns, equipment operation schedules, and energy usage differences in functional areas) to precisely characterize the dynamic changes in building electricity load and improve model prediction accuracy.
- (2) Addressing the issue of abnormal load fluctuations under extreme weather conditions, research could focus on adaptive improvement methods for outliers. By enhancing the model's adaptability to extreme data, reliable prediction performance can be maintained even under special operating conditions.
- (3) Considering the real-time control requirements of building energy systems, future work could explore lightweight model design. While ensuring prediction accuracy, the computational complexity and inference latency of the model could be reduced to enable deployment on edge devices and achieve real-time prediction applications.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Natural Science Foundation of China (52577115), Education Research Project for Young and Middle-Aged Teachers of Fujian Provincial Education Department (JAT251119) and Startup Fund for Advanced Talents of Putian University (2023133).

Author Contributions: The authors confirm contribution to the paper as follows: methodology and writing—original draft: Chuan Lin; investigation, validation, visualization: Weixian Chen; formal analysis: Guangtao Hao. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, C. Lin, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. *Renew Sustain Energy Rev.* 2018;81:1192–205. doi:10.1016/j.rser.2017.04.095.
2. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Trans Smart Grid.* 2019;10(1):841–51. doi:10.1109/tsg.2017.2753802.
3. Zhou M, Wang L, Hu F, Zhu Z, Zhang Q, Kong W, et al. ISSA-LSTM: a new data-driven method of heat load forecasting for building air conditioning. *Energy Build.* 2024;321(2):114698. doi:10.1016/j.enbuild.2024.114698.
4. Chen Z, Yang Y, Jiang C, Chen Y, Yu H, Zhou C, et al. Enhanced industrial heat load forecasting in district networks via a multi-scale fusion ensemble deep learning. *Expert Syst Appl.* 2025;272(1):126783. doi:10.1016/j.eswa.2025.126783.
5. Semmelmann L, Hertel M, Kircher KJ, Mikut R, Hagenmeyer V, Weinhardt C. The impact of heat pumps on day-ahead energy community load forecasting. *Appl Energy.* 2024;368(3):123364. doi:10.1016/j.apenergy.2024.123364.
6. Huang Y, Zhao Y, Wang Z, Liu X, Fu Y. Sparse dynamic graph learning for district heat load forecasting. *Appl Energy.* 2024;371(3):123685. doi:10.1016/j.apenergy.2024.123685.

7. Shi Z, Zheng R, Shen R, Yang D, Wang G, Liu Y, et al. Building heating load forecasting based on the theory of transient heat transfer and deep learning. *Energy Build.* 2024;313(12):114290. doi:10.1016/j.enbuild.2024.114290.
8. Huang Y, Zhao Y, Wang Z, Liu X, Liu H, Fu Y. Explainable district heat load forecasting with active deep learning. *Appl Energy.* 2023;350:121753. doi:10.1016/j.apenergy.2023.121753.
9. Li Y, Zhu N, Hou Y. A novel hybrid model for building heat load forecasting based on multivariate Empirical modal decomposition. *Build Environ.* 2023;237(12):110317. doi:10.1016/j.buildenv.2023.110317.
10. Bergsteinsson HG, Møller JK, Nystrup P, Pálsson ÓP, Guericke D, Madsen H. Heat load forecasting using adaptive temporal hierarchies. *Appl Energy.* 2021;292(8):116872. doi:10.1016/j.apenergy.2021.116872.
11. Somu N, Raman MRG, Ramamritham K. A deep learning framework for building energy consumption forecast. *Renew Sustain Energy Rev.* 2021;137:110591. doi:10.1016/j.rser.2020.110591.
12. Xiao JW, Liu P, Fang H, Liu XK, Wang YW. Short-term residential load forecasting with baseline-refinement profiles and bi-attention mechanism. *IEEE Trans Smart Grid.* 2024;15(1):1052–62. doi:10.1109/tsg.2023.3290598.
13. Bourdeau M, Zhai XQ, Nefzaoui E, Guo X, Chatellier P. Modeling and forecasting building energy consumption: a review of data-driven techniques. *Sustain Cities Soc.* 2019;48(10):101533. doi:10.1016/j.scs.2019.101533.
14. Chou JS, Ngo NT. Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. *Appl Energy.* 2016;177:751–70. doi:10.1016/j.apenergy.2016.05.074.
15. Fumo N, Rafe Biswas MA. Regression analysis for prediction of residential energy consumption. *Renew Sustain Energy Rev.* 2015;47(8):332–43. doi:10.1016/j.rser.2015.03.035.
16. Chen Y, Xu P, Chu Y, Li W, Wu Y, Ni L, et al. Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings. *Appl Energy.* 2017;195:659–70. doi:10.1016/j.apenergy.2017.03.034.
17. Deng S, Dong X, Tao L, Wang J, He Y, Yue D. Multi-type load forecasting model based on random forest and density clustering with the influence of noise and load patterns. *Energy.* 2024;307(12):132635. doi:10.1016/j.energy.2024.132635.
18. Han F, Pu T, Li M, Kong Z, Chen H, Wang X. Short-term forecasting of individual residential load based on deep learning and K-means clustering. *CSEE J Power Energy Syst.* 2021;7:261–9. doi:10.17775/CSEEJPES.2020.01250.
19. Huang X, Han Y, Yan J, Zhou X. Hybrid forecasting model of building cooling load based on EMD-LSTM-Markov algorithm. *Energy Build.* 2024;321(3):114670. doi:10.1016/j.enbuild.2024.114670.
20. Tang X, Xia Y, Jiang L, Xiong D, Wang L, Wang Y. Dynamic adaptive hierarchical TCN driven by IHOA-VMD optimization for short term load forecasting. *Energy.* 2025;335(4):138074. doi:10.1016/j.energy.2025.138074.
21. Liu H, Li Z, Li C, Shao L, Li J. Research and application of short-term load forecasting based on CEEMDAN-LSTM modeling. *Energy Rep.* 2024;12(7):2144–55. doi:10.1016/j.egyr.2024.08.035.
22. Li K, Huang W, Hu G, Li J. Ultra-short term power load forecasting based on CEEMDAN-SE and LSTM neural network. *Energy Build.* 2023;279:112666. doi:10.1016/j.enbuild.2022.112666.
23. Imani M. Electrical load-temperature CNN for residential load forecasting. *Energy.* 2021;227(1):120480. doi:10.1016/j.energy.2021.120480.
24. Fan P, Wang D, Wang W, Zhang X, Sun Y. A novel multi-energy load forecasting method based on building flexibility feature recognition technology and multi-task learning model integrating LSTM. *Energy.* 2024;308(3):132976. doi:10.1016/j.energy.2024.132976.
25. Rubasinghe O, Zhang X, Chau TK, Chow YH, Fernando T, Iu HH. A novel sequence to sequence data modelling based CNN-LSTM algorithm for three years ahead monthly peak load forecasting. *IEEE Trans Power Syst.* 2024;39(1):1932–47. doi:10.1109/tpwrs.2023.3271325.
26. Deka P, Cordeiro-Costas M, Pérez-Orozco R, Chabiński M, Szłęk A. Novel NSGA-II optimized LSTM and GRU models for short-term forecasting of residential heating load. *Energy Build.* 2025;344:115999. doi:10.1016/j.enbuild.2025.115999.
27. Wang J, Chen X, Zhang F, Chen F, Xin Y. Building load forecasting using deep neural network with efficient feature fusion. *J Mod Power Syst Clean Energy.* 2021;9(1):160–9. doi:10.35833/mpce.2020.000321.

28. Hua H, Liu M, Li Y, Deng S, Wang Q. An ensemble framework for short-term load forecasting based on parallel CNN and GRU with improved ResNet. *Electr Power Syst Res.* 2023;216(3):109057. doi:10.1016/j.epsr.2022.109057.
29. Chen W, Rong F, Lin C. A multi-energy loads forecasting model based on dual attention mechanism and multi-scale hierarchical residual network with gated recurrent unit. *Energy.* 2025;320(9):134975. doi:10.1016/j.energy.2025.134975.
30. Li C, Shi J. A novel CNN-LSTM-based forecasting model for household electricity load by merging mode decomposition, self-attention and autoencoder. *Energy.* 2025;330(1):136883. doi:10.1016/j.energy.2025.136883.
31. Hu Z, Ma R. Adaptive event-triggered tracking control via switching functions. *Automatica.* 2026;185(4):112813. doi:10.1016/j.automatica.2025.112813.
32. Hu Z, Liu S, Luo W, Wu L. Intrusion-detector-dependent distributed economic model predictive control for load frequency regulation with PEVs under cyber attacks. *IEEE Trans Circuits Syst I.* 2021;68(9):3857–68. doi:10.1109/tcsi.2021.3089770.