Electronic Research Archive

http://www.aimspress.com/journal/ERA

ERA, 32(3): 2160–2179.
DOI: 10.3934/era.2024098
Received: 02 December 2023
Revised: 30 January 2024
Accepted: 06 February 2024
Published: 14 March 2024

Research article

TCN-Attention-BIGRU: Building energy modelling based on attention mechanisms and temporal convolutional networks

Yi Deng[1,2,*], Zhanpeng Yue[1], Ziyi Wu[1], Yitong Li[3] and Yifei Wang[4]

[1] School of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China
[2] State Key Laboratory of New Textile Materials and Advanced Processing Technologies, Wuhan Textile University, Wuhan 430200, China
[3] School of Electronic and Information Engineering, Hankou University, Wuhan 430212, China
[4] School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China

* **Correspondence:** Email: ydeng@wtu.edu.cn.

Abstract: Accurate and effective building energy consumption prediction is an important basis for carrying out energy-saving evaluation and the main basis for building energy-saving optimization design. However, due to the influence of environmental and human factors, energy consumption prediction is often inaccurate. Therefore, this paper presents a building energy consumption prediction model based on an attention mechanism, time convolutional neural (TCN) network fusion, and a bidirectional gated cycle unit (BIGRU). First, t-distributed stochastic neighbor embedding (T-SNE) was used to preprocess the data and extract the key features, and then a BIGRU was employed to acquire past and future data while capturing immediate connections. Then, to catch the long-term dependence, the dataset was partitioned into the TCN network, and the extended sequence was transformed into several short sequences. Consequently, the gradient explosion or vanishing problem is mitigated when the BIGRU handles lengthy sequences while reducing the spatial complexity. Second, the self-attention mechanism was introduced to enhance the model's capability to address data periodicity. The proposed model is superior to the other four models in accuracy, with an mean absolute error of 0.023, an mean-square error of 0.029, and an coefficient of determination of 0.979. Experimental results indicate that T-SNE can significantly improve the model performance, and the accuracy of predictions can be improved by the attention mechanism and the TCN network.

Keywords: building energy consumption; feature selection; gated cycle unit; attention mechanism; time convolutional neural network; prediction

1. Introduction

In modern energy management systems, the conservation and use of building energy have received great attention. Many large public buildings can predict future short- or medium-term building energy consumption data by establishing energy-saving systems to save resources and minimize business costs [1]. Energy consumption in large commercial buildings is influenced by many factors, such as the weather and national holidays, so the accuracy of energy prediction requires further studies [2]. According to a literature review, the common methods for constructing energy prediction models are classified as physical models [3], statistical models, and machine learning models [4–5]. Creating a physical model for analyzing building energy consumption is a computer simulation technique [6] that requires collecting the building shape, materials, insulation, lighting, ventilation, and other factors, often using CAD, Sketchup, Revit, and other modeling software, but can also use professional building energy analysis software such as Energy Plus to complete the building model. The accuracy of the physical model's predictions is greatly influenced by input characteristic parameters, such as temperature, humidity, dew point, and visibility [7]. These parameters are difficult to obtain, resulting in sub-optimal physical models. Statistical models obtain the relationship between characteristic parameters and influencing factors by modeling them using statistics. Models based on statistical principles, such as linear regression [8] and multivariate linear regression, are computationally fast and suitable for exploring the linear relationships between variables, but cannot deal with non-linear relationships, are sensitive to outliers, and do not capture complex energy consumption patterns well. Hosseinzadeh et al. [9] proposed support vector machines (SVMs), which can handle complex non-linear relationships, have strong generalization ability, and are robust to high-dimensional data and outliers, but their parameter tuning is more complex, computational overhead is greater, and training time for large datasets can be longer than that of other methods. The rapid development of artificial intelligence has increased the accuracy of machine learning methods, which are gradually replacing applied research in statistical methods. Machine learning models can be categorized as single models, like the random forest algorithm proposed by Georganos et al. [10]. Liu et al. [11–13] proposed new methods in the field of deep learning, which includes natural language processing and image recognition. This algorithm elaborates on the random forest's construction process, highlights the importance of feature selection, and showcases its wide applicability to classification and regression problems. Bentéjac et al. [14] introduced the gradient boosting algorithm, which delineates the gradient boosting tree's principles, the training procedure, and hyper-parameter tuning. Additionally, it covers this algorithm's applications to classification, regression, and ranking problems. Nevertheless, because these methods employ pre-determined nonlinearities, they might not be able to identify the nonlinearities in different datasets, resulting in low robustness. To address the inadequacy of simple model prediction methods, complex model prediction methods have been suggested for energy consumption prediction. Kiruthika and Thaila [15] formulated a building energy prediction model, long short-term memory (LSTM)-SVM, which exhibits good prediction performance. Li and Fan [16] achieved a strong predictive performance by adopting the particle swarm optimization (PSO) technique to optimize the back propagation (BP) model. However, statistical methods have a restricted capability to extract time series features and cannot accommodate the nonlinear and unstable features of building energy consumption well. Deep learning methods, particularly recurrent neural networks (RNNs) [4] and their variations, are becoming increasingly prevalent in energy consumption prediction. Variants of RNNs, such as LSTM and gated recurrent unit (GRU) [10], address RNN's long-term dependency problem and are suitable for energy consumption prediction applications. In [17], Hewamalage et al. [17] introduced a new network architecture that combines LSTM networks with convolutional neural networks (CNNs). The architecture employs a novel LSTM-CNN approach [18], and experiments demonstrated the method's superior ability. Priyadarshini and Cotton [18] proposed a hybrid deep learning model based on feature extraction and CNNs and experimentally demonstrated its predictive performance. In [19], Aslam et al. [19] proposed a model combining LSTM and GRU, with LSTM reducing the dimensionality of the features and GRU capturing the relationships between data in the time series. The model has high prediction accuracy. Li et al. [20] proposed a short-term wind power forecasting model based on deep learning and error correction. The bi-directional long short-term memory (BILSTM) model is used for prediction, and the principal component analysis algorithm is used to construct and continuously correct the error model. The effectiveness and applicability of the model were verified through experiments. In [21], Niu et al. [21] proposed a wind power prediction model based on the attention mechanism and the transformer model. In the model, the attention mechanism determines key information, and the model could also predict well. However, in real engineering, variable factors can affect the prediction results, so introducing the attention mechanism into the energy consumption prediction model can enhance the flexibility, accuracy, and interpretability of the model, especially when dealing with complex and variable energy consumption patterns. The mechanism can be used to help the model understand and predict the factors that influence the consumption of energy and thus provide valuable guidance for energy management and conservation [7]. Bagal et al. [22] proposed an energy consumption forecasting model based on transformers, where an attention mechanism is introduced to weight building characteristics and climate data to increase the forecasting accuracy. Yuan et al. [23] used the CNN model based on the attention mechanism to predict building system consumption over time for a university in Shanghai. The results show that the attention mechanism improves the data processing capability of CNN.

Predicting energy consumption in buildings has received great research attention, but current machine learning methods can only extract some of the features due to the complexity of meteorological features and the variability of the number of workers in the building. Therefore, as an innovation in the construction of the predictive model, this study combines data denoising and smoothing to simplify the basic parameters of the model and fully extract the features of the data. T-SNE is used to denoise the data [24]. To reduce the influence of abnormal data on prediction accuracy, visualization and comparison experiments are carried out with three types of dimensionality reduction. Meanwhile, the bidirectional gated cycle unit (BIGRU) model is selected as the base model [25]. This simplified structure helps improve operational efficiency. The BIGRU model can be applied to the prediction of building energy consumption data on a large scale. BIGRU is used for the appropriate extraction of time series features to highlight the influence of different time series nodes on energy consumption, an attention mechanism is introduced to increase the weighting of key time steps, and high-level time series features are further extracted by a temporal convolution network (TCN) layer. The combined effect of these elements is a significant improvement in the model's operating efficiency, stability, and accuracy. The model can help provide statistical support for safe building operation.

The main contributions of this paper can be summarized as follows:

1. The use of TCN with BIGRU can capture features on different time scales and is thus proposed

for the first time.

2. The introduction of the attention mechanism into the above model improves the problem of dealing with nonlinear relationships and multi-feature sequences.

3. The experimental results show that our model has a higher prediction accuracy than the four mainstream models used for comparison.

The remainder of the paper is structured as follows: Section 2 introduces related work. Section 3 describes the functionality of each module. Section 4 presents the experimental results and analysis, and Section 5 presents the conclusion.

2. Related work

2.1. Building energy consumption forecast

As the main body of public buildings, office buildings have a high energy-saving potential and low transformation costs. Accurate and effective building energy consumption prediction is an important basis for the selection of energy-saving retrofit programs for existing office buildings and for the optimal energy-saving design of new buildings.

Currently, physical and data models, which respectively use thermodynamic principles and machine learning techniques, are the most common methods of forecasting energy use. Given that physical models are time-consuming and require detailed building information and environmental parameters (e.g., building construction details, operation schedules, physical parameters) [3], the lack of accurate input data often leads to poor energy consumption simulation results in practical applications. With the increase of data volume and arithmetic power, data models are gradually being developed from shallow machine learning to deep learning. Deep learning extracts and learns features directly from data through multiple network layers and has a strong model representation capability. Compared with traditional machine learning, the prediction accuracy of deep learning models increases with the training data. Thus, many researchers have made use of software simulation, statistical analysis, neural networks, and SVMs to forecast the energy usage of buildings and obtained particular outcomes. However, with the increasing number of energy and building types, and given the characteristics of large-scale public building energy consumption, the conventional methods for predicting energy consumption in buildings are increasingly facing issues, such as low prediction accuracy, non-real-time results, and an extremely long prediction cycle, preventing them from meeting the actual needs. Researchers have been attempting to address these problems. Liu et al. [26–29] proposed a new approach about perceptual networks and feature learning in the industrial field.

Given the above issues, this paper designs a set of building energy consumption prediction models, the model has better prediction accuracy than the traditional prediction model and meets the realistic demand.

2.2. Data denoising

Data denoising is a commonly used method for reducing data dimensionality. The basic idea of data denoising is to identify and remove noise from the original data to retain useful signals and patterns. In deep learning, data denoising can be categorized into reduction based on dimensionality methods and dimensionality reduction based on regression data denoising, which reduce the effect of noise by projecting the data into a lower dimensional space [30] to allow the extraction of the main features and patterns in the data. These methods are usually used to deal with high-dimensional data to facilitate analysis and modeling. PCA (principal component analysis) was proposed by Hewage [31], achieving dimensionality reduction by projecting the data in the direction that maximizes its variance. In this process, dimensions with a small variance are usually considered noise and can be removed by selecting fewer principal components. However, dimensionality reduction methods usually map the data from a high-dimensional space to a low-dimensional space, which may result in some information loss. Although this method helps remove noise, it may also ignore some important features in the data.

Data denoising based on regression methods is the process of identifying and removing noise by building a regression model to obtain cleaner data. The basic idea of this method is to identify outliers that may be due to noise by fitting a regression model to the data and then analyzing the residuals of the model (the difference between the predicted and actual values). The isolation forest method proposed by Gao et al. [32] is an integration-based regression method for identifying outliers. The method constructs a model by randomly partitioning the data and then evaluates the data point outliers to identify noise. The thresholding of the residuals may then lead to excessive denoising.

2.3. Attention mechanism

Considering the correlating cycles within the dataset, this study also uses a time-specific attention mechanism [18]; attention can be used to improve the understanding of how similar characteristics relate to each other and the extraction of characteristics using different weights for each characteristic. In particular, the attention layer trains its own data and constantly updates parameters to obtain the weight of attention. Using the output sequence generated by the TCN network as input into the attention module [33], this study obtains a set of implicit attention with weights, which can adapt to the frequency of the data set.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right). \tag{1}$$

where Q denotes the proposed model, which is used to actively compute the degree of similarity with other tokens. K denotes the model proposed from the token, and the relationship information with other tokens is used to calculate the degree of similarity with other tokens. V denotes the importance of the current token.

2.4. BIGRU model

LSTM and GRU are extensions of RNNs [34], which cannot solve the problem of increasing sequence length. In response to this problem, researchers made a related upgrade and then proposed an upgraded version of the RNN LSTM and a high-level version of the GRU, facilitating the solution for the ladder loss and explosion problems. The GRU model has few parameters, which are easy to train and have low calculation complexity, and in the RNN-LSTM model, the memory and input gates become new gates, and they and the reset gates form a new unit. BIGRU, which is also known as the two-way GRU network, is a variant of GRU which realizes positive and reverse two-way propagation so that the output layer carries out a reverse propagation on the basis of the forward propagation. Each training sequence is presented in both forward and backward directions to two distinct hidden layers.

These layers are then linked to the same output layer, which subsequently contains complete information on past and future occurrences for each point in the input sequence. Compared with the one-way GRU model, BIGRU converges quickly and does not easily overfit. BIGRU's output depends on the double weight of the forward state and the backward state influence, increasing the accuracy of the final output result. GRU combines the current node input x^t with the state h^{t-1} transmitted down from the previous node to derive the output y^t of the current node and the hidden state h^t passed to the next node. Parameter passing and updating equations within the network as shown in Eqs (2)–(5):

$$r^t = \sigma(w^{rx}x^t + w^{rh}h^{t-1} + b^r). \tag{2}$$

$$z^t = \sigma(w^{zx}x^t + w^{zh}h^{t-1} + b^z). \tag{3}$$

$$h' = \tan h\,(w^{xh}x^t + r^t \odot w^{hh}h^{t-1}). \tag{4}$$

$$h' = (1 - z^t) \odot h^{t-1} + h' \odot z^t. \tag{5}$$

Where σ is the sigmoid function, r is the gating control for reset, z is the gating control for update, and h refers to the candidate hiding state. w^{rx}, w^{rh} etc. are the weight matrices, b^r, b^z etc. are the biases. $\odot$ is the Hadamard product, i.e., multiplying the elements in one matrix corresponding elements in the other.

3. Proposed TCN-Attention-BIGRU method

This paper proceeds as follows: division of the collected energy consumption data into data and validation sets, data prediction using the TCN-Attention-BiGRU framework, and presentation of the prediction results using three metrics. This section explains how these methods work.

3.1. TAB model

A model for predicting the energy consumption of a building using TCN-BIGRU and an attention mechanism is described in this paper. The structure and workflow of the model are illustrated in Figure 1. The model is primarily partitioned into the Input, Data Preprocessing, TCN, BIGRU, Attention, and Output layers. With the use of historical power and real-time meteorological and mathematical data for each property, the TCN layer extracts the features. The BIGRU and attention layers learn the internal power change law of the proposed features to realize the prediction function. Finally, the prediction results are obtained through the output layer.

The dimensions of each batch in the proposed model's input layer are 8 and 3, that is, the time step is 8 and the number of sample features is 3. The TCN layer and the GRU are combined to process the input data in both the forward and backward directions of the bidirectional layer. The features learned by the two one-way GRUs are then concatenated to produce a set of vectors that serve as the input to the attention layer. The TCN layer expands the perceived field size of the convolution by setting multiple extension and causal convolution layers and multiplying the weight vector with the output vector of the BIGRU layer to obtain the output of the attention layer. The model parameters are shown in Table 1.

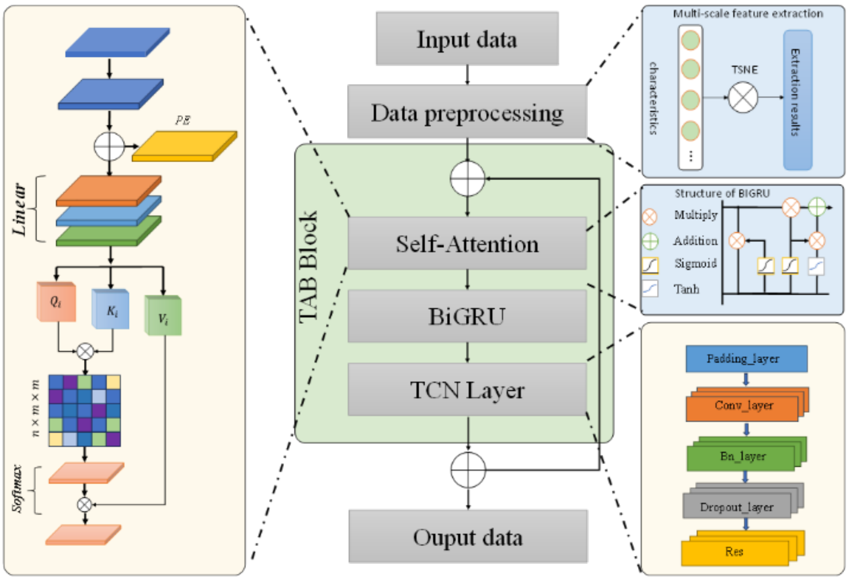

Figure 1. Structure of TCN-Attention-BIGRU.

Table 1. TCN-Attention-BIGRU parameter.

Hyperparameter	Value
TCN filters	64
Activation	ReLU
BIGRU units	64
Activation	sigmoid
Attention activation	softmax
Epoch	60
Loss function	mean-square error (MSE)

3.2. *Attention-TCN*

The attention mechanism performs selective learning of information by acquiring the importance of the information. Inspired by [23], we design TCN-Attention, an energy efficient predictive attention module. The TCN-Attention structure is shown in Figure 2.

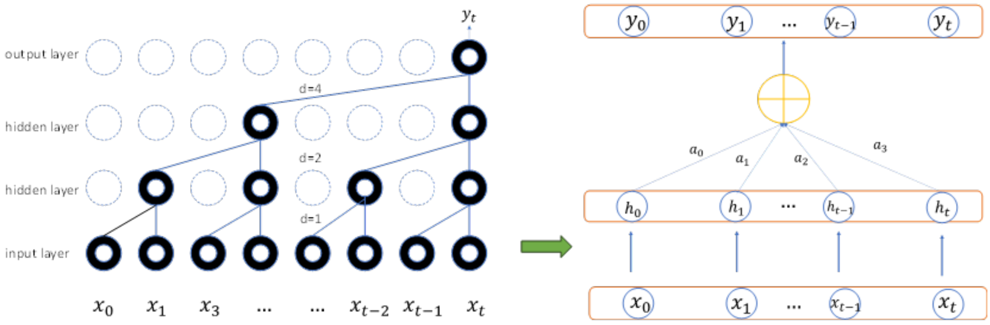

Figure 2. Structure of the TCN-Attention module.

This module can effectively avoid the problem of gradient vanishing or exploding faced by RNNs, has the advantages of parallel computation and low memory consumption, and can effectively filter a few crucial features from a larger set of features and increase their significance while decreasing the importance of non-critical features to emphasize the influence of key ones. x_t (t∈ $[0,n]$) denotes the input of the TCN network, h_t(t∈ $[0,n]$) corresponds to the hidden layer output obtained from each input through the TCN, a_t(t∈ $[0,n]$) is the attention weight of the attention mechanism for the hidden layer output of the TCN, and y_t(t∈ $[0,n]$) is the output layer in which the attention mechanism is introduced. The weight coefficients of the attention mechanism can be expressed as follows:

$$e_t = u\tanh(wh_t + b), \tag{6}$$

$$a_t = \frac{\exp(e_t)}{\sum_{j=0}^{t} e_j}, \tag{7}$$

$$y_t = \sum_{t=0}^{n} a_t h_t. \tag{8}$$

3.3. Feature selection

Input feature selection is especially critical for improving the performance of the prediction model and reducing the computational cost. In this study, the Pearson correlation coefficient (PCC) is used to reflect the necessity of feature extraction. The PCC is calculated as

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{9}$$

where x_i, y_i denote the variables， and $\bar{x}$, $\bar{y}$ denote their mean values.

$|\rho| \geq 0.8$: strong correlation; $0.6 \leq |\rho| < 0.8$: relatively strong correlation; $0.4 \leq |\rho| < 0.6$: moderate correlation; $0.2 \leq |\rho| < 0.4$: weak correlation; and when $0 \leq |\rho| < 0.2$, it is irrelevant

Figure 3 shows that the absolute PCC value for visibility [34], dew point, and humidity is less than 0.4. Therefore, these features should be discarded. We analyzed the reasons for the weak relationship between these three features and the total energy consumption. For the teaching building, the relationship is mostly dynamic, so visibility and others do not have a strong correlation with energy consumption, such as the temperature and wind speed, which is strongly related to the number of people in the teaching building and therefore has a strong correlation. In summary, the energy consumption characteristics of the building must be extracted.

The confusion matrix shows that energy consumption is more closely related to atmospheric pressure, wind speed, and temperature, and analyzing the following correlations between them can be concluded:

The correlation coefficient between the energy consumption value and atmospheric pressure is 0.3988, the correlation coefficient between the energy consumption value and wind speed is 0.9038, and the correlation coefficient between the energy consumption value and temperature is −0.7511, which all pass the 0.05 confidence test.

In Figure 3, the Pearson coefficient between the energy consumption value and wind speed reaches 0.9. This is because, in a high wind speed environment, the building is subjected to a large airflow impact, which leads to an increase in the temperature loss of the building's exterior wall, thus increasing the energy consumption of the building. In addition, air circulation inside the building is also enhanced in high wind speed environments, leading to uneven heat distribution inside the building and increasing the energy consumption of the heating and cooling systems. On the other hand, the low wind speed environment decreases the temperature loss of the building facade, which reduces the energy consumption of the building. Therefore, there is a positive correlation between the energy consumption of the building and the outdoor wind speed, i.e., the higher the wind speed is, the higher the energy consumption of the building is; the lower the wind speed is, the lower the energy consumption of the building is.

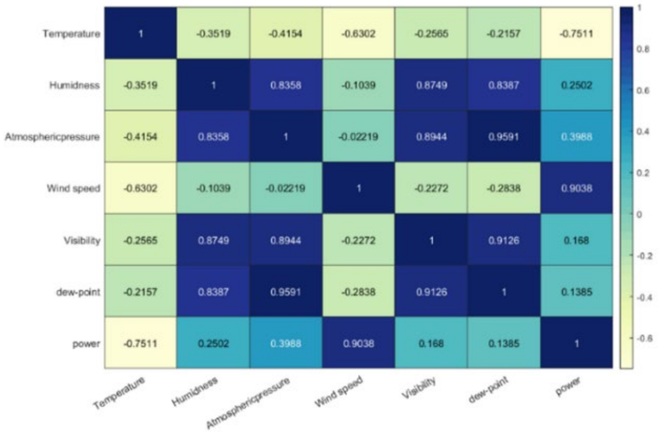

Figure 3. PCCs of different variables in construction.

As shown in Table 2, these methods improve the performance of the model but have some shortcomings.

Table 2. Comparison of data denoising methods.

Method	Data downgrading	Feature retention
PCA	√	×
LDA	√	×
Gaussian process regression	×	√
Isolated forest	×	√
T-SNE (proposed method)	√	√

In this paper, by applying T-SNE, the data can be downscaled to lower dimensions so that the features that play an important role in describing the variation in the data are easy to identify.

According to the T-SNE principle, the main parameters that affect the final result include the projection dimension and confusion degree. First, the relevant parameters are determined. To determine the optimal parameters, Kullback-Leibler Divergence (KL) divergence is used as a measurement index. The smaller the KL divergence is, the better the dimensionality reduction effect is. Given the interaction between the confusion degree and the projection dimension, the exhaustion method is used to select the best parameters. First, the KL divergence between projection dimensions 1 and 11 is calculated by fixing the confusion degree. The comparative analysis shows that KL divergence is smallest when the projection dimension is 2, and the KL divergence gradually increases with the increase of the projection dimension. According to the preferred projection dimension, the norm limit of the confusion degree is set as 5–50, and the KL divergence value is calculated. The KL divergence increases in a wavy manner with the increase of the confusion degree, and the optimal confusion degree is determined to be 5. Second, after determining the optimal dimension and confusion degree, the variance of each sample point is calculated using the dichotomy σ_i, and the joint probability matrix of high-dimensional space $P = (Pij)_{348 \times 348}$ is calculated using Eq 12. The random initialization of the projected data is expressed as $X^* = (x_i^*)_{348\, x\, 2}$. According to Eqs 13 and 14, the low-dimensional joint probability and the KL divergence value are calculated, respectively. Finally, the gradient descent method is used to perform iterative operations on X, and the output result is the projection data in the low-dimensional space.

The specific steps of denoising using the T-SNE method are as follows:

(1) T-SNE converts the distance into a joint probability distribution with symmetric characteristics through the Gaussian distribution to reflect the similarity of the high-dimensional spatial data.

$$P_{j|i} = \frac{\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)}{\sum_{k\neq l} exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)}, \tag{10}$$

where $x_i x_j$ represents two points, and $P_{j|i}$ denotes the probability that x_j is its nearest neighbor point when the center is x_i.

(2) To overcome the defect of SNE data point crowding, a T-distribution with 1 degree of freedom is constructed to convert the distance into a probability distribution to evaluate the similarity of the low-dimensional data.

$$q_{ij} = \frac{\left(1+\|y_i-y_2\|^2\right)^{-1}}{\sum_{k\neq l}(1+\|y_i-y_2\|^2)^{-1}}, \tag{11}$$

where y_i, y_2 are the points corresponding to the mappings of x_i, x_j into the lower dimensional space.

(3) The M-divergence of the two distributions are optimized to construct the following objective function. The gradient descent formula is

$$C = M(P \parallel Q) = \sum_{i,j} P_{i,j} \log\frac{p_{ji}}{q_{ji}}, \tag{12}$$

$$\frac{\delta C}{\delta y_i} = 4\sum_j \left(p_{ji} - q_{ji}\right)(y_i - y_2)(1 + \parallel y_i - y_2 \parallel^2)^{-1}. \tag{13}$$

(4) Two spatial M-divergence targets are minimized by using the gradient descent method, and the low-dimensional visualization parameters are calculated at the same time to complete the functions of data dimensionality reduction and visualization.

In this study, PCA, linear discriminate analysis (LDA), and independent component correlation algorithm (ICA) are compared with the proposed method [28,29]. The principal elements of PCA, LDA, and ICA are independent of each other in the low-dimensional space. Consequently, data points that are uncorrelated in the high-dimensional space but are very near each other may be far from each other in the low-dimensional space. However, in the T-SNE algorithm, the correlation coefficient of the projected samples does not need to be 0. Therefore, the neighborhood of the data in the high-dimensional space can be maintained in the low-dimensional space so that the samples in the same cluster are near each other, the samples in different clusters are far from each other, and the local structural features of the original data are preserved. Therefore, from the data visualization perspective, the T-SNE algorithm has a better visualization effect than the other models. We normalize the data to obtain 2280 training sets, 1200 validation sets, and 2300 test sets for each feature parameter, followed by dimensionality reduction clustering analysis of the processed feature parameters, as shown in Figure 4.

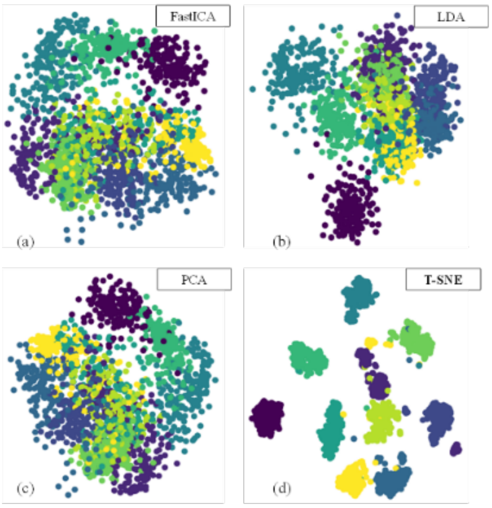

Figure 4. Visualization results of (a) FastICA, (b) LDA, (c) PCA, and (d) T-SNE.

4. Experimental results

The entire experimental process can be further divided into three stages: preprocessing, training, and testing. In the preprocessing phase, the original energy dataset is normalized and partitioned into 8:1:1 for training, validation, and testing. The training phase includes T-SNE downscaling and BIGRU neural network training. The testing phase uses different evaluation metrics for performance comparison. In this section, the performance of the proposed TAB method is compared with existing cutting-edge techniques in the literature, including the following:

I. GRU.

II. Autoregressive Integrated Moving Average model (ARIMA).

III. Generative Adversarial Network (GAN).

Additionally, there are also more advanced models:

I. Transformer.

II.Timenet

4.1. Model evaluation indicators

Neural network optimization minimizes the loss function, which can reflect the quality of the model and provide the direction of optimization. The weight of the model is updated by the gradient descent method, that is, the inverse direction obtained by the derivation of the weight is updated according to the loss function. Different optimization tasks have different loss functions. When the target is a binary or multi-classification task, the cross-entropy loss function is generally used. When the goal is a regression problem, the mean square error and the mean absolute error are generally used as loss functions. Considering that the task in this study is a prediction problem, which is also equivalent to the effect of regression, the MSE and the mean absolute error (MAE) can be considered as the loss functions of this study. The MAE is obtained using Eq (9), while the MSE expression is obtained in Eq (14).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \widehat{y_i}|. \tag{14}$$

where y_i is the real value, $\widehat{y_i}$ is the predicted value, and m is the number of samples.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2. \tag{15}$$

Similarly, y_i, $\widehat{y_i}$ has the same meaning as above.

The coefficient of determination (R^2) reflects the accuracy of the model in fitting the data; generally, the R^2 range is 0 to 1. The closer the value is to 1, the stronger the explanation ability of the equation variables to y, and the better the fit of the model to the data.

$$R^2 = 1 - \frac{\sum_i(\widehat{y_i} - y_i)^2}{\sum_i(\bar{y_i} - y_i)^2} \times 100\%. \tag{16}$$

4.2. Data description

The raw energy consumption data named “CN-Data” was collected from a university building in Hubei Province. The building is a teaching building, which is a building used for conducting teaching activities. The main function of the teaching building is to provide classrooms for conducting classroom teaching. The influencing factors have a strong impact on the energy consumption values of the teaching building, such as the use of the building and the number of personnel, external environmental conditions such as climate and air temperature, and the architectural design and structure of the building, etc. Therefore, the energy consumption data of the teaching building is used as the dataset for this experiment, and the energy consumption data such as humidity, barometric pressure, wind speed, visibility, and dew point are used as the input variables of the model, and the transformation of the multiple time steps is not done by downward sampling, but rather the accumulation of energy consumption is calculated on the basis of time intervals. The capacity of the dataset for each building is shown in Table 3.

Table 3. Size of the datasets from the three buildings (with 5 min intervals).

	Total	**Training**	**Validation set**	**Testing set**
Building 1	22478	2280	1200	2300
Building 2	4540	980	1200	2300
Building 3	3723	680	600	1150

Figure 5 shows that the necessity of feature extraction is verified due to the large differences in energy consumption between office and non-office hours in the academic building and variation between different features.

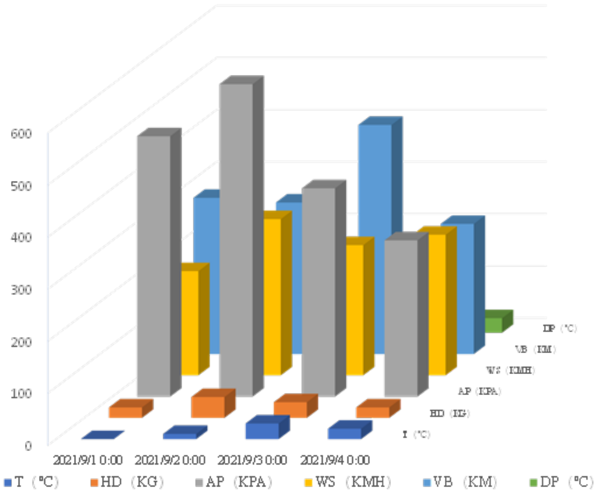

Figure 5. Changes in variables for four consecutive days.

4.3. Data normalization

Given that the six energy consumption factors, such as temperature, humidity, visibility, and dew point, are of different orders of magnitude, the data should be standardized to eliminate the dimensional influence between the quantitative indicators. The data standardization method is shown in Eq (17):

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{17}$$

where $\mathrm{x_{max}}$ is the maximum value in a specific dimension of sample data, $\mathrm{x_{min}}$ is the minimum value in a specific dimension of sample data, and x is the sample data in this dimension, thus achieving the standardization of a single dimension. In fact, the calculation of multiple dimensions standardizes the data in multiple dimensions to achieve the effect of normalization.

4.4. Results

4.4.1. Model evaluation

Tables 4–6 show the statistical results for the three data sets.

Table 4. Results of the experiment evaluated on the dataset of Academic Building 1.

	MAE			MSE			R^2		
Interval (min)	15	40	60	15	40	60	15	40	60
GRU	0.026	0.089	0.155	0.042	0.119	0.211	0.312	0.283	0.224
Arima	0.018	0.071	0.163	0.037	0.121	0.262	0.297	0.274	0.252
GAN	0.014	0.051	0.127	0.035	0.108	0.189	0.315	0.301	0.297
Transformer	0.031	0.063	0.134	0.025	0.068	0.089	0.641	0.743	0.521
TimeNet	0.024	0.044	0.079	0.043	0.089	0.088	0.543	0.691	0.771
TCN-Attention-BiGRU	0.008	0.017	0.023	0.005	0.020	0.029	0.991	0.982	0.979

Table 5. Results of the experiment evaluated on the dataset of Academic Building 2.

	MAE			MSE			R^2		
Interval (min)	15	40	60	15	40	60	15	40	60
GRU	0.023	0.089	0.153	0.042	0.119	0.211	0.312	0.283	0.223
Arima	0.015	0.073	0.163	0.037	0.152	0.272	0.242	0.231	0.291
GAN	0.016	0.056	0.126	0.038	0.108	0.188	0.315	0.352	0.599
Transformer	0.072	0.051	0.136	0.022	0.047	0.086	0.643	0.742	0.513
TimeNet	0.041	0.024	0.067	0.023	0.074	0.073	0.144	0.699	0.761
TCN-Attention-BiGRU	0.008	0.017	0.021	0.004	0.021	0.028	0.981	0.972	0.963

Table 6. Results of the experiment evaluated on the dataset of Academic Building 3.

	MAE			MSE			R^2		
Interval (min)	15	40	60	15	40	60	15	40	60
GRU	0.023	0.081	0.147	0.043	0.118	0.213	0.332	0.281	0.227
Arima	0.021	0.072	0.165	0.039	0.153	0.245	0.296	0.274	0.276
GAN	0.013	0.081	0.141	0.031	0.111	0.181	0.314	0.304	0.294
Transformer	0.053	0.045	0.105	0.033	0.045	0.076	0.443	0.546	0.908
TimeNet	0.013	0.063	0.133	0.043	0.087	0.177	0.667	0.797	0.908
TCN-Attention-BiGRU	0.006	0.014	0.021	0.007	0.024	0.029	0.922	0.986	0.977

The MAE, MSE, and R^2 values of the compared methods are shown in the previous tables. Among the compared methods, the proposed method has the lowest error rate. The experimental results indicate that the proposed method has more accurate results than the traditional prediction methods.

In Figure 6, the fitted curves of the real and predicted values are shown. The closer the curves are, the more accurate the prediction is, and the last one is the prediction curve of this model.

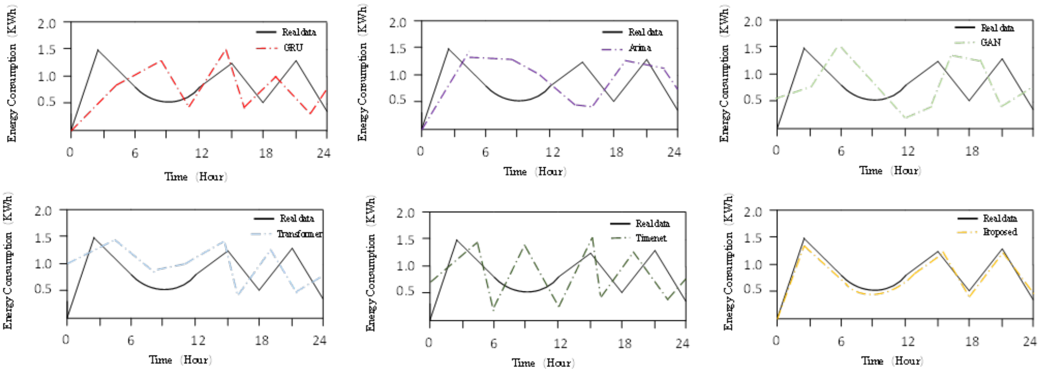

Figure 6. Forecasting results for CN-Data energy consumption data of Building 1.

4.4.2. Parameter sensitivity analysis

In order to assess the degree of importance of each input parameter in the dataset, the input parameters atmospheric pressure A, wind speed W, and temperature T were multiplied by coefficients of variation of 1.2 and 1.5, respectively, for the prediction set of data, and then predicted, and the resulting performance evaluation metrics are shown in Table 7.

Table 7. Parameters sensitivity analysis.

variation parameter	Coefficient of variation	MAE (m)	MSE (m)	R^2 (%)
T	1.2	2.76	2.31	2
	1.5	2.80	2.14	0
W	1.2	0..92	2.19	81
	1.5	0.68	1.19	30
A	1.2	0.80	0.73	83

As can be seen from Table 7, according to the order of MAE, the parameter temperature T is more sensitive and has a greater degree of influence on the energy consumption value, while the parameters atmospheric pressure A and wind speed W have a smaller influence on the energy consumption value. It can be seen that temperature is the main influencing factor of the energy consumption value, followed by wind speed, and atmospheric pressure has the least influence on energy consumption value. This is also in line with common sense that, in cold climates, buildings need to be provided with heating to maintain a comfortable temperature. Lower outdoor temperatures will result in a drop in the internal temperature of the building, requiring more heat from the heating system. This will increase the energy consumption of the heating system. In hot climates, buildings need to use air conditioning systems to lower indoor temperatures. Higher outdoor temperatures will result in higher temperatures inside the building, requiring more cooling from the air conditioning system. This will increase the energy consumption of the air conditioning system.

4.4.3. Model comparison

All the comparative results are displayed in Figure 6, in which the real energy consumption data is shown with black bars, and the red color indicates the prediction results of this model. Traditional machine learning methods, such as GRU and GAN, have a large shortfall in prediction accuracy, and the combinatorial model addresses this, but it remains inferior to the proposed model.

4.4.4. Effect of time intervals on the model

The operational data of the teaching building possesses unique characteristics. Identifying the working day time of the teaching staff as 7:00 to 18:00, in Figure 7(a) is the fitting curve between the predicted and actual values of energy consumption during working hours, and in Figure 7(b) is the fitting curve between the predicted and actual values of energy consumption during non-working hours, from which it can be seen that the prediction effect during working hours is very good and very close to the real value, and the prediction effect during non-working days is very poor, which is mainly because, during the daytime, people carry out their work, classes, or other activities inside the building, which require energy consumption such as lighting, heating, or cooling, etc. At night, on the other hand, personnel activities are reduced and only a few necessary equipment and lighting need to be used, thus energy consumption is lower. This difference leads to different accuracy in predicting daytime and nighttime energy consumption. Also, daytime is usually sunny and warmer, requiring more air conditioning and ventilation to maintain comfortable indoor temperatures. In contrast, nighttime temperatures drop, demand is lower, and energy consumption decreases accordingly. This change in the external environment affects the prediction results of building energy consumption. Building energy consumption involves the combined effects of several factors, such as building characteristics, equipment effectiveness, and personnel activities. The complexity of the energy consumption model may result in different accuracy of energy consumption prediction for specific time periods.

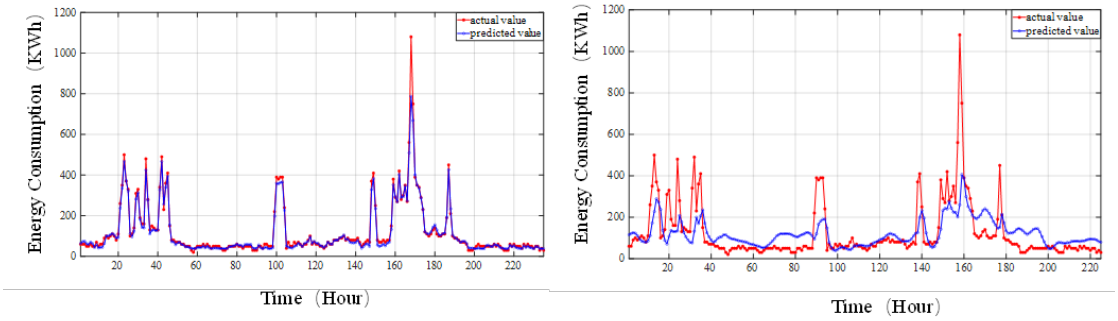

Figure 7. Model predictions for working time: (a) working time datasets, (b) non-working time datasets.

4.4.5. Parameter and time window width evaluation

In this model, the number of hidden layers and the optimizer choice are important considerations. The model is iterated using different parameters, and the best parameters are selected, as shown in Figure 8. The optimal number of nodes in this model is 64, and the optimal optimizer is the Adam (Adaptive Moment Estimation) optimizer [35].

In the prediction of building energy consumption, the length of the window has a great influence on the extraction of key information and reduction of computation. Therefore, the appropriate sliding window size should be selected. This study verifies the model accuracy when the sliding window length is 6, 12, 24, and 36, respectively, and Figure 8 shows that the prediction effect is the best when the sliding window length is 24.

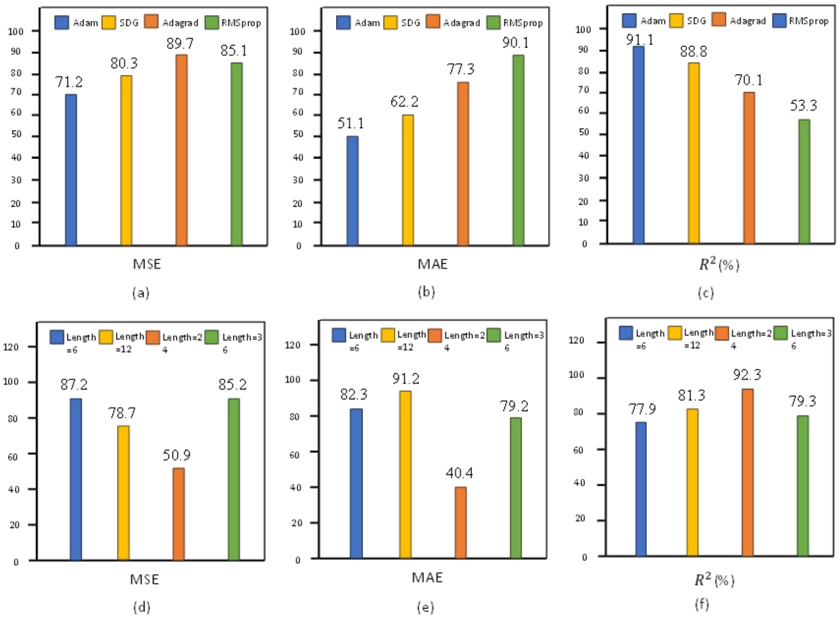

Figure 8. Model optimizer and hidden layer evaluation.

5. Conclusions

In this study, a BIGRU system based on the attention mechanism is proposed for building energy consumption prediction.

The model extracts the main features in the data through T-SNE and adds the processed data as inputs to the model, introduces the attention mechanism module, focuses on the information that has an important impact on the energy consumption prediction results by assigning different weights to different features, reduces the loss of historical data, and reduces the importance of the non-critical features by using the TCN to emphasize the impact of the critical features, and the main conclusions are: The T-SNE data dimensionality reduction method proposed in this paper can extract more relevant data and improve the performance of the model compared to methods such as PCA. The superiority of the method is verified by comparing it with six classical models. The MAE and MSE of the method are reduced by at least 1.5 and 2.1%, respectively, and the R^2 is improved by 10.8%. Compared with combined models such as GRU, the method reduces the MAE and MSE by at least 1.3 and 1.7%, respectively, and improves the R^2 by 7.8%. The best model prediction performance was achieved when the number of sliding windows was 24 and the optimizer was Adam. In addition, the effect of time period on prediction was discussed, and the effect of hyperparameters and sliding windows on model performance was compared with GRU, GAN, and ARIMA models. When the dataset was divided into workdays and non-workdays, the predictions for workdays showed a high goodness of fit, while the predictions for non-workdays showed a very low goodness of fit, which suggests that there is no point in predicting the energy consumption of a building at night. The T-SNE method was used for dimensionality reduction and was compared with the other three methods.

The research results in this paper have higher accuracy than current advanced prediction models, provide a basis for future research on related topics, and are of practical significance: the use of the building energy consumption prediction model in this paper helps to set and achieve energy efficiency and emission reduction targets, and timely adjustment of building energy consumption schedules in regions with large temperature variations helps to cope with extreme climatic conditions and mitigate impacts on the energy system.

This paper proposes a new idea for building energy consumption prediction. However, there are some limitations in this paper. This paper found that different network structures have an impact on energy consumption prediction, but there is no in-depth research on the mechanism of the impact, and so in the future it can be introduced into the transformation model for research attempting to use different network structures, model parameter adjustments, or other modeling techniques to improve the accuracy and generalization ability of the model. Building energy consumption prediction can obtain information from multiple data sources, such as energy consumption data, weather data, and people activity data. Future research could explore how to effectively fuse these multimodal data to improve the predictive power and accuracy of the model. Liu et al.'s [36,37] proposal about attention networks and recommendation algorithms sheds a lot of light on this paper. These are very promising directions for further research.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. D. Li, M. Qiu, J. Jiang, S. Yang, The application of an optimized fractional order accumulated grey model with variable parameters in the total energy consumption of Jiangsu Province and the consumption level of Chinese residents, *Electron. Res. Arch.*, **30** (2022), 798–812. https://doi.org/10.3934/era.2022042
2. M. Aydin, N. I. Mahmudov, H. Aktuğlu, E. Baytunç, M. S. Atamert, On a study of the representation of solutions of a ψ-Caputo fractional differential equations with a single delay, *Electron. Res. Arch.*, **30** (2022), 1016–1034. https://doi.org/10.3934/era.2022053
3. C. Ohajunwa, C. Caiseda, P. Seshaiyer, Computational modeling, analysis and simulation for lockdown dynamics of COVID-19 and domestic violence, *Electron. Res. Arch.*, **30** (2022), 2446–2464. https://doi.org/10.3934/era.2022125
4. J. Zheng, Y. Li, Machine learning model of tax arrears prediction based on knowledge graph, *Electron. Res. Arch.*, **31** (2023), 4057–4076. https://doi.org/10.3934/era.2023206
5. X. Shen, P. Raksincharoensak, Statistical models of near-accident event and pedestrian behavior at non-signalized intersections, *J. Appl. Stat.*, **49** (2022), 4028–4048. https://doi.org/10.1080/02664763.2021.1962263
6. Q. Li, D. Huang, S. Pei, J. Qiao, M. Wang, Using physical model experiments for hazards assessment of rainfall-induced debris landslides, *J. Earth Sci.*, **32** (2021), 1113–1128. https://doi.org/10.1007/s12583-020-1398-3
7. L. Xu, F. Chen, F. Ding, A. Alsaedi, T. Hayat, Hierarchical recursive signal modeling for multifrequency signals based on discrete measured data, *Int. J. Adapt. Control Signal Process.*, **35** (2021), 676–693. https://doi.org/10.1002/acs.3221
8. D. Alita, A. D. Putra, D. Darwis, Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation, *Indones. J. Comput. Cybern. Syst.*, **15** (2021), 295–306. https://doi.org/10.22146/ijccs.65586
9. M. Hosseinzadeh, A. M. Rahmani, B. Vo, M. Bidaki, M. Masdari, M. Zangakani, Improving security using SVM-based anomaly detection: issues and challenges, *Soft Comput.*, **25** (2021), 3195–3223. https://doi.org/10.1007/s00500-020-05373-x
10. S. Georganos, T. Grippa, A. N. Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, et al., Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling, *Geocarto Int.*, **36** (2021), 121–136. https://doi.org/10.1080/10106049.2019.1595177
11. H. Liu, T. Liu, Y. Chen, Z. Zhang, Y. Li, EHPE: Skeleton cues-based gaussian coordinate encoding for efficient human pose estimation, *IEEE Trans. Multimedia*, (2022), 1–12. https://doi.org/10.1109/TMM.2022.3197364
12. H. Liu, C. Zhang, Y. Deng, T. Liu, Z. Zhang, Y. Li, Orientation cues-aware facial relationship representation for head pose estimation via transformer, *IEEE Trans. Image Process.*, **32** (2023), 6289–6302. https://doi.org/10.1109/TIP.2023.3331309

13. H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, Z. Zhang, et al., Trans-IFC: Invariant cues aware feature concentration learning for efficient fine-grained bird image classification, *IEEE Trans. Multimedia*, (2023), 1–14. https://doi.org/10.1109/TMM.2023.3238548
14. C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, *Artif. Intell. Rev.*, **54** (2021), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5
15. N. S. Kiruthika, D. G. Thaila, Dynamic light weight recommendation system for social networking analysis using a hybrid LSTM-SVM classifier algorithm, *Opt. Mem. Neural Networks*, **31** (2022), 59–75. https://doi.org/10.3103/S1060992X2201009X
16. S. Li, Z. Fan, Evaluation of urban green space landscape planning scheme based on PSO-BP neural network model, *Alexandria Eng. J.*, **61** (2022), 7141–7153. https://doi.org/10.1016/j.aej.2021.12.057
17. H. Hewamalage, C. Bergmeir, K. Bandara, Recurrent neural networks for time series forecasting: Current status and future directions, *Int. J. Forecast.*, **37** (2021), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008
18. I. Priyadarshini, C. Cotton, A novel LSTM-CNN-grid search-based deep neural network for sentiment analysis, *J. Supercomput.*, **77** (2021), 13911–13932. https://doi.org/10.1007/s11227-021-03838-w
19. N. Aslam, F. Rustam, E. Lee, P. B. Washington, I. Ashraf, Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model, *IEEE Access*, **10** (2022), 39313–39324. https://doi.org/10.1109/ACCESS.2022.3165621
20. M. Li, D. Xu, J. Geng, W. Hong, A ship motion forecasting approach based on empirical mode decomposition method hybrid deep learning network and quantum butterfly optimization algorithm, *Nonlinear Dyn.*, **107** (2022), 2447–2467. https://doi.org/10.1007/s11071-021-07139-y
21. Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, *Neurocomputing*, **452** (2021), 48–62. https://doi.org/10.1016/j.neucom.2021.03.091
22. V. Bagal, R. Aggarwal, P. K. Vinod, U. D. Priyakumar, MolGPT: Molecular generation using a transformer-decoder model, *J. Chem. Inf. Model.*, **62** (2021), 2064–2076. https://doi.org/10.1021/acs.jcim.1c00600
23. Y. Yuan, Z. Chen, Z. Wang, Y. Sun, Y. Chen, Attention mechanism-based transfer learning model for day-ahead energy demand forecasting of shopping mall buildings, *Energy*, **270** (2023), 126878. https://doi.org/10.1016/j.energy.2023.126878
24. D. Kobak, G. C. Linderman, Initialization is critical for preserving global data structure in both t-SNE and UMAP, *Nat. Biotechnol.*, **39** (2021), 156–157. https://doi.org/10.1038/s41587-020-00809-z
25. T. Ahmad, H. Chen, Y. Guo, J. Wang, A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review, *Energy Build.*, **165** (2018), 301–320. https://doi.org/10.1016/j.enbuild.2018.01.017
26. T. Liu, H. Liu, B. Yang, Z. Zhang, Limb direction cues-aware network for flexible human pose estimation in industrial behavioral biometrics systems, *IEEE Trans. Ind. Inf.*, (2023), 1–11. https://doi.org/10.1109/TII.2023.3266366
27. H. Liu, T. Liu, Z. Zhang, A. K. Sanga, B. Yang, Y. Li, ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction, *IEEE Trans. Ind. Inf.*, **18** (2022), 7107–7117. https://doi.org/10.1109/TII.2022.3143605

28. H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, J. Wang, MFDNET: Collaborative poses perception and matrix fisher distribution for head pose estimation, *IEEE Trans. Multimedia*, **24** (2021), 2449–2460. https://doi.org/10.1109/TMM.2021.3081873
29. H. Liu, C. Zheng, D. Li, X. Shen, K. Lin, J. Wang, et al., EDMF: Efficient deep matrix factorization with review feature learning for industrial recommender system, *IEEE Trans. Ind. Inf.*, **18** (2022), 4361–4371. https://doi.org/10.1109/TII.2021.3128240
30. D. Liu, W. Wang, X. Wang, C. Wang, J. Pei, W. Chen, Posts seismic data denoising based on 3-D convolutional neural network, *IEEE Trans. Geosci. Remote Sens.*, **58** (2020), 1598–1629. https://doi.org/10.1109/TGRS.2019.2947149
31. A. Daffertshofer, C. J. C. Lamoth, O. G. Meijer, P. J. Beek, PCA in studying coordination and variability: a tutorial, *Clin. Biomech.*, **19** (2004): 415–428. https://doi.org/10.1016/j.clinbiomech.2004.01.005
32. L. Gao, J. Gao, J. Li, A. Plaza, L. Zhuang, X. Sun, et al., Multiple algorithm integration based on ant colony optimization for endmember extraction from hyperspectral imagery, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **8** (2014), 2569–2582. https://doi.org/10.1109/JSTARS.2014.2371615
33. P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, et al., Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station, *Soft Comput.*, **24** (2020), 16453–16482. https://doi.org/10.1007/s00500-020-04954-0
34. Y. Yu, L. You, D. Liu, W. Hollinshead, Y. J. Tang, F. Zhang, Development of Syne sp. PCC 6803 as a phototrophic cell factory, *Mar. Drugs*, **11** (2013), 2894–2916. https://doi.org/10.3390/md11082894
35. A. K. Shahade, K. H. Walse, V. M. Thakare, Deep learning approach-based hybrid fine-tuned Smith algorithm with Adam optimiser for multilingual opinion mining, *Int. J. Comput. Appl. Technol.*, **73** (2023), 50–65. https://doi.org/10.1504/IJCAT.2023.134080
36. H. Liu, C. Zheng, D. Li, Z. Zhang, K. Lin, X. Shen, et al., Multi-perspective social recommendation method with graph representation learning, *Neurocomputing*, **468** (2022), 469–481. https://doi.org/10.1016/j.neucom.2021.10.050
37. B. A. Draper, K. Baek, M. S. Bartlett, J. R. Beveridge, Recognizing faces with PCA and ICA, *Comput. Vision Image Understanding*, **91** (2003), 115–137. https://doi.org/10.1016/S1077-3142(03)00077-8