

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/168503/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Jiang, Ben, Li, Yu, Rezgui, Yacine , Zhang, Chengyu, Wang, Peng and Zhao, Tianyi 2024. Multi-source domain generalization deep neural network model for predicting energy consumption in multiple office buildings. *Energy* 299 , 131467. 10.1016/j.energy.2024.131467

Publishers page: <http://dx.doi.org/10.1016/j.energy.2024.131467>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

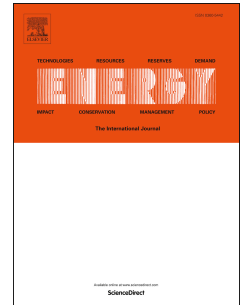
This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Journal Pre-proof

Multi-source Domain Generalization Deep Neural Network Model for Predicting Energy Consumption in Multiple Office Buildings

Ben Jiang, Yu Li, Yacine Rezgui, Chengyu Zhang, Peng Wang, Tianyi Zhao



PII: S0360-5442(24)01240-4

DOI: <https://doi.org/10.1016/j.energy.2024.131467>

Reference: EGY 131467

To appear in: *Energy*

Received Date: 23 December 2023

Revised Date: 2 April 2024

Accepted Date: 27 April 2024

Please cite this article as: Jiang B, Li Y, Rezgui Y, Zhang C, Wang P, Zhao T, Multi-source Domain Generalization Deep Neural Network Model for Predicting Energy Consumption in Multiple Office Buildings, *Energy*, <https://doi.org/10.1016/j.energy.2024.131467>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 Published by Elsevier Ltd.

Multi-source Domain Generalization Deep Neural Network Model for Predicting Energy Consumption in Multiple Office Buildings

Ben Jiang^a, Yu Li^b, Yacine Rezgui^c, Chengyu Zhang^a, Peng Wang^a, Tianyi Zhao^{a,*}

a. Institute of Building Energy, Dalian University of Technology, Dalian, China.

b. College of Environment Science and Engineering, Donghua University, Shanghai, China.

c. School of Engineering, Cardiff University, Cardiff, UK.

* Corresponding author: zhaotianyi@dlut.edu.cn

Abstract

The effective prediction of building energy consumption can be used to optimize building operating modes and reduce the overall energy consumption and carbon emission of the building. To predict the energy consumption of the same type of building in the same area, it is often necessary to train separate prediction models for different buildings, which is usually costly computationally in order to get better prediction results. Therefore, this study will combine deep neural networks and multi-source domain generalization in an encoder-decoder architecture to train a model that can be directly applied to predict the energy consumption of multiple buildings in same type. In order to validate the predictive effectiveness of the model, it will be tested on real office building energy consumption dataset and different comparative experiments will be designed on the source and target domains. The results show that the constructed multi-source domain generalization model is able to accurately predict the energy consumption trend of different buildings one hour in advance. It also has some energy consumption prediction ability for the unknow training set of buildings.

Keywords:

Energy consumption prediction; Deep neural network; Multi-source domain generalization; Encoder and decoder architecture; Office buildings

1.Introduction

Building consumes a large amount of energy, about 40 percent of total global energy and more than 30 percent of greenhouse gas emissions[1-4]. Therefore, it is crucial to provide adequate energy demand for equipment in buildings as well as to optimize the mode of operation in order to achieve energy savings. And energy consumption prediction has received a lot of attention from researchers as a basis for energy management tasks. Its results can be used to optimize building operations and reduce the overall carbon emissions of the building.[5, 6].

Data driven approaches are becoming increasingly popular in energy consumption prediction studies within the building sector. Data driven approaches are favored because they do not require large amounts of physical information about the building, and are skilled at describing complex relationships between data variables. As a result, these approaches have been widely utilized in predicting trends in building energy consumption[7, 8]. At this stage, the primary data driven algorithms include statistical regression and machine learning algorithms. Machine learning algorithms are widely used for their advantages of dealing with nonlinear data, such as Support

Vector Machines(SVM)[9], Random Forests[10], and Artificial Neural Networks (ANN)[11]. Traditional ANN models have independent input and output, lacking methods to process time series data. To address this, Recurrent Neural Networks (RNN) incorporating temporal correlation have been developed to solve such issues[12]. Through the continuous improvement of RNN and the development of deep neural networks, more effective prediction models have been invented and applied to the task of predicting building energy consumption, such as Long Short-Term Memory(LSTM)[13], Gate Recurrent Unit(GRU)[14] and the neural network combining Convolutional Neural Networks(CNN) and LSTM (CNNLSTM)[15]. However, during the model training process, equal attention is typically given to different input data, potentially leading to bias in the prediction results[16]. The emergence of Transformer[17] has introduced the multihead attention mechanism, which has been widely used in some energy consumption prediction tasks and has yielded favorable prediction results[18-20].

In many energy consumption prediction tasks, favorable results can be achieved by employing appropriate machine learning methods. However, the traditional approach for predicting energy consumption in buildings of the same type within a specific area involves training a separate prediction model for each building. In order to get better predictions, it is often necessary to consume large amounts of computational resources. Furthermore, models trained for one task may not be applicable to other similar tasks, even when sufficient data is available[21-23]. To address this challenge, Transfer Learning (TL) has become a promising solution. As a machine learning method, TL divides the datasets into source domain and target domain. Some of the algorithms, such as TL for fine-tuning strategies[24, 25], Domain Adaptation Network[26, 27] and Domain Adversarial Neural Network[28, 29] have a wide range of applications in computer vision, image categorization, linguistic analysis, and sentiment analysis[30, 31]. Domain Adaptation Network is a commonly used method in TL to effectively overcoming the problem of data distribution discrepancy, and has been widely used in the field of energy consumption prediction in recent years[8, 32-34]. Ribeiro[35] and others, using a time series adaptive transfer learning approach to remove temporal effects and perform a prediction task on the target domain improved energy consumption prediction results in schools by 12 percent. Fang[36] and others, improved the prediction results of the target domain by identifying domain invariant features between multiple source and target domains. This is mainly achieved through a multi-source transfer learning framework that integrates neural networks and similarity metrics. But they also found that the prediction performance of the model degraded when there are too many source domains. Lu[37] and others, used Dynamic Time Warping (DTW) for source domain selection and combined it with the target domain to train a multi-source domain energy prediction model on a test set. The prediction results were compared with those of single source domain and found that the mean absolute percentage error (MAPE) was optimized by 6.88%-15.37%. At this stage, domain adaptive modeling in the field of energy consumption prediction through multi-source or single target domains has been studied to some extent. However, a generalization model that can correspond to multiple target domains in the same area with the same type of building complex may be more applicable. This would allow for the development of prediction models with good results for both target and source domain buildings on energy consumption prediction tasks using a small amount of training resources.

Domain Generalization (DG) may offer a solution to these challenges. DG is a commonly used method in TL, requiring a given number of source domains to construct a model for the target task that can make accurate predictions about unknown data distributions. Domain generalization can be

viewed as an extension the problem of Domain Adaptive[38]. Methods based on DG have made significant progress in other areas such as natural language processing and computer vision[30, 39, 40]. These methods can be broadly categorized into three groups, namely Data Augmentation, Learning Invariant Representation for Domain and Learning Strategies. Among these, Learning Invariant Representation for Domain is a common framework for enhancing the generalization capabilities of models[41, 42]. In time series forecasting tasks, the phenomenon of Temporal Covariate Shift (TCS) often occurs, and the forecasting model can be improved by decomposing the unknown data into the least similar sequences and training a generalization model to extract the time invariant features[43]. Learning more timeseries invariant properties might help to enhance the prediction of source and target domains in the training of energy consumption prediction for multi-source generalization models.

The objective of this study is to train an energy consumption prediction generalization model on a deep neural network using data from multiple source domains under the boundary conditions of the same building type in the same area. The model can show good predictive performance on both the target and source domains without accessing the target domain data during training. This reduces the cost of modeling each building independently. For some buildings of the same type with unknown training set or less data, it is hoped that the model can also be used directly for the task of energy consumption prediction for such buildings with some prediction ability. To achieve this goal, the study sets a one hour ahead building energy consumption prediction task and used real datasets to conduct experiments. The DTW algorithm is used for source domain data selection. To enhance the feature extraction capability of the model, a deep neural network is constructed using a combination of CNN, LSTM, and Attention within an encoder-decoder architecture. A multi-source generalizaion model is trained on the source domain using Temporal Distribution Characterization (TDC) to quantify time series similarity. Finally, performance of the model in this task is compared with other models on the test sets of both the source and target domains. Tree-structured Parzen Estimator(TPE) Bayesian optimization is employed to optimize hyperparameters for the deep neural network model and the comparative models to ensure fairness in the comparative experiments. The study is divided into five parts. The second part mainly describes the construction of the deep neural network architecture, the selection of the source domain, and the construction of the multi-source domain generalization model. The third part explains the data sources, experimental design, and data preprocessing processes. The fourth part discusses the predictive results, and the study concludes with the main findings is the last part.

2 Research Methodology

2.1 Deep Neural Network Architecture for Energy Consumption Prediction

2.1.1 Convolutional Neural Networks

CNN is a commonly used algorithm for data feature extraction and has been widely applied to various time series prediction tasks[44, 45]. It is adept at processing data with a distinct grid topology and can handle input data in different formats, including 1D, 2D, and nD [46]. A typical convolutional neural network architecture comprises several key components, such as a convolutional layer, a pooling layer, a fully connected layer, an output layer, and various activation functions. The continuous processing of input signals in CNN results in a closer representation of

the initial signals to the output signals, thereby enhancing the capability of model to handle the task. In this study, 1D CNN will serve as the initial feature extraction module for the input data, with ReLU being utilized as the activation function in the CNN.

2.1.2 Long Short-Term Memory Neural Networks

LSTM neural network is shown in Fig.1. The LSTM has implemented more refined internal processing units and has been applied to a large number of learning tasks related to time series.

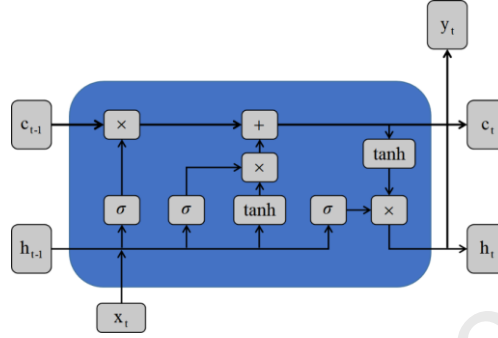


Fig.1. LSTM structure diagram

The LSTM consists of three main parts, the forget gate, the input gate and the output gate. Each input contains two hidden states, the initial cell state (c) and the initial hidden state (h) to represent the output of the previous cell, which are initialized with a 0 tensor and x_t represents the input of the current cell. First, when computing the next hidden state c_t , a decision needs to be made about the information f_t to be discarded through the forget gate. When the information from the previous step is introduced, the information g_t introduced in the current step and the information i_t needs to control the flow into the neural network is computed from the current input tensor x_t and the implied state h_{t-1} of the previous step combined with the activation function. Finally, the new implied state h_t is computed through the output gate to obtain the output o_t of the cell. The relevant formula is as follows:

$$f_t = \sigma(W_{xf}x_t + b_f + W_{hf}h_{t-1} + b_{hf}) \quad (1)$$

$$i_t = \sigma(W_{xi}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \quad (2)$$

$$g_t = \tanh(W_{xg}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \quad (4)$$

Where σ is the Sigmoid activation function, $\tanh$ is the Tanh activation function, W_x is the linear transformation parameter of x , and W_h is the linear transformation parameter of h . The new state information c_t and the final result y_t can be calculated by combining the implied state information from the previous step:

$$c_t = f_t \times c_{t-1} + i_t \times g_t \quad (5)$$

$$y_t = h_t = o_t \times \tanh c_t \quad (6)$$

In general, the LSTM controls the output state by selection as a way to remember the context and is more suitable for time series prediction tasks.

2.1.3 Attention Mechanism

Attention mechanism has been used in the development of deep neural networks for a variety of tasks at this stage[47, 48]. In particular, the multi-head attention mechanism mentioned in Transformer, scale factor and multihead mechanism can capture correlations between multiple sequences[17]. The ordinary self-attention mechanism requires three values, the query tensor (Q), the key-value tensor (K), and the value tensor (V).

$$Query = QW_i^Q \quad (7)$$

$$Key = KW_i^K \quad (8)$$

$$Value = VW_i^V \quad (9)$$

Where i is the ordinal number of the attention weight and W is the weight matrix, calculated as:

$$Attention(Q, K, V) = softmax\left(\frac{Query \cdot Key^T}{\sqrt{d_i}}\right) Value \quad (10)$$

Where d_i is the feature value of the i th attentional weight to enhance the stability of the algorithm. The formula for the multihead attention mechanism is as follows.

$$Multihead = Concat(head_1, head_2, \dots, head_h)W \quad (11)$$

The multihead attention mechanism uses multiple weights to weight the results of the calculation and then stitches all the results together.

2.1.4 Deep Neural Network Algorithm Architecture

In this section, a deep neural network architecture (CNN-AttLSTM) will be constructed to more accurately predict future energy consumption trends from historical energy consumption data for the target building. The flexibility of the encoder-decoder architecture makes them a common choice as the primary architecture for deep neural networks in time series prediction tasks[49-51]. The encoder and decoder in this study mainly include CNN, LSTM, multihead Attention, and fully connected layers (Linear).

The architecture of the deep neural network is shown in Fig.2. During the operation process, the first input x_i ($i=1, 2, 3, \dots, n$) passes through each layer of the network to compute the encoder output y_j ($j=1, 2, 3, \dots, t$). This output is then input into the decoder operation to compute the prediction value y'_i . The error between the predicted value y'_i and the actual data y_i is calculated and used to update the network parameters.

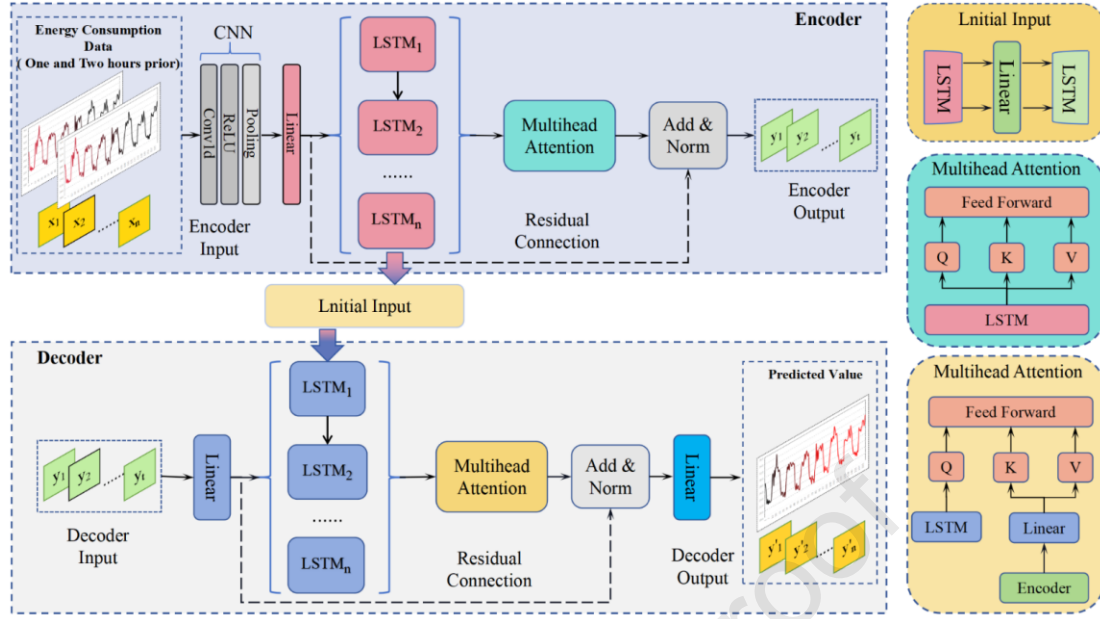


Fig.2. Deep Neural Network Algorithm Architecture

For error calculation, the Mean Square Error (MSE) loss function will be employed, with the formula as follows, where n represents the number of samples.

$$MSE = \frac{\left(\sum_{i=1}^{i=n} (y_i - y_i')^2 \right)}{n} \quad (12)$$

The network architecture also contains a number of optimizations for computing. That includes using the LSTM output of the encoder to initialize operations on the LSTM of the decoder, employing the output of the encoder as the K and V values for computation in the multihead attention module of the decoder following the example of the Transformer, and ensuring stable model training through the addition of the residual network.

2.2 Multisource Domain Generalization Algorithm Architecture

2.2.1 DTW Algorithm

In the process of TL, the more similarities between the target and source domains, the easier to transfer knowledge from the source domain to the target domain, thus minimizing the risk of negative transfer. Various methods have been utilized in recent years to compute the similarity of two sets of data, such as the Pearson correlation coefficient, Spearman correlation coefficient, Maximum Mean Discrepancy (MMD), and DTW. Among these, DTW has shown significant efficacy in certain prediction tasks for deep neural networks[52-54]. DTW employs dynamic programming to calculate the distance between two time series and is widely applied in the field of time series similarity matching[55].

During the DTW calculation process, an $m \times n$ matrix is constructed for two time series x_i ($i=1, 2, \dots, m$) and x_j ($j=1, 2, \dots, n$), with the distance $d(x_i, x_j)$ corresponding to each point being computed and filled in, typically using the Euclidean distance. The Dynamic Programming (DP) algorithm is then utilized to find a path H in this matrix grid, starting at $(i, j) = (1, 1)$ and ending at $(i, j) = (m, n)$. When passing through the point (i, j) in the grid, the previous point can only be $(i-1, j)$, $(i-1, j-1)$ and $(i, j-1)$. The DP algorithm is employed to compute the optimal H while satisfying the aforementioned

conditions, thereby minimizing the cumulative distance h on H . A smaller h indicates a higher similarity between the two time series.

In this study, the distances $h_{z,r}$ ($r=1, 2, \dots, k, z \in k \& z \neq r$), mean h'_z and variance h''_z from each building to the remaining $k-1$ buildings are calculated on the training set to screen the source domain buildings. Initially, the three buildings with the smallest h'_z are filtered out, and then the buildings with a greater degree of fluctuation in their h value are further eliminated using h''_z . Finally, the remaining two buildings are selected as source domain buildings to ensure a higher level of similarity to the other buildings in terms of overall energy consumption data. This also serves to reduce the number of buildings in the target domain whose calculated distances from the source domain buildings deviate significantly from the overall mean.

2.2.2 Temporal Distribution Characterization

As show in Fig.3, the data distribution $P(x)$ of a time series may be various at different time stages, and the distribution for unknown test data may also be different from the training data. This will greatly affect the predictive stability of the prediction model.

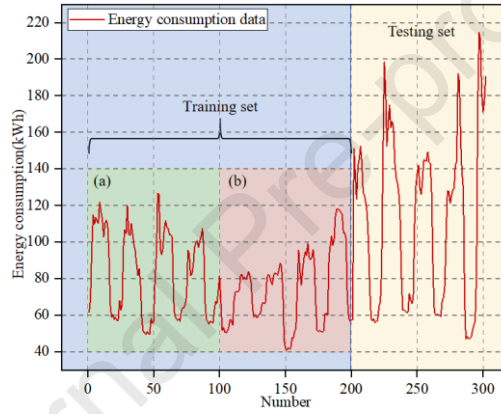


Fig.3. Example of data distribution for test and training sets of energy consumption data

Adaptive RNNs (AdaRNN) was invented in order to get time series prediction models with better results in prediction[43]. The problem of time series distribution change is defined as Temporal Covariate Shift (TCS), and AdaRNN can find the data segments with the largest distribution difference in the training datasets through the Temporal Distribution Characterization (TDC) module, which realizes better generalization and prediction ability than the traditional RNN by reducing the distribution difference of these data segments. The main purpose of the TDC module is to quantize and classify the continuous data in a time series into sequences with the least similar K segment distribution by using the following formula.

$$\max_{0 < K \leq K_0} \max_{n_1, \dots, n_K} \frac{1}{K} \sum_{1 \leq i \neq j \leq K} d(D_i, D_j) \quad (13)$$

$$\text{s. t. } \forall i, \Delta_1 < |D_i| < \Delta_2; \sum_i |D_i| = n \quad (14)$$

Where Δ_1 , Δ_2 and K_0 are predetermined values to avoid meaningless solutions. d is a distance metric that can be an arbitrary distance function. In this study, Deep Correlation Alignment (CORAL) will be used as d for computation. CORAL is defined as the distance of the second-order statistic (covariance) of the samples from the two distributions, and its distance expression is computed by the following formula.

$$d_{\text{coral}}(h_s, h_t) = \frac{1}{4q^2} \|C_s - C_t\|_F^2 \quad (15)$$

Where q is the dimensionality of the features, C_s and C_t are the covariance matrices of the two distributions respectively. The time series is uniformly partitioned by the TDC algorithm into N parts, where each part is the smallest unit not divisible anymore. Then, an appropriate value of K will be determined in the range of $[2, N]$. For example when $K=2$, the goal is then to find a C point to maximize the distance d .

In this study, the energy consumption data of the previous one and two hour are used to predict the fluctuation trend of energy consumption in the coming hour. So the training set will be segmented by using the TDC algorithm based on the inputs of the model. In order to optimize the computational process and to extract information useful for all building prediction tasks, the following operations will be performed in this paper. Each source domain training set was first divided into six parts on average based on the input parameters. Secondly using TDC algorithm the source domains with smaller DTW average are divided into three parts and others are divided into two parts. Finally the divided data is used for the training of multi-source generalization model.

2.2.3 Multi-source Generalization Model Architecture

The architecture of the multi-source generalization model (TDC-CNN-AttLSTM) training is shown in Fig.4. Firstly, the v source domain buildings S_v for training the generalization model are filtered by DTW calculation. The different source domain data are divided using the TDC algorithm and then the segmented data are entered sequentially into the encoder architecture for training. The output of the encoder is passed through a linear layer, then the intra-domain and inter-domain distance losses are computed without repetition, and the computed results are summed up to give the total distance loss. The output of the encoder is also sent to the decoder for training, and the error loss between the predicted and true values is calculated. The total network is trained using the energy prediction multi-source domain generalization model, with the objective of minimizing both the distance loss and error loss.

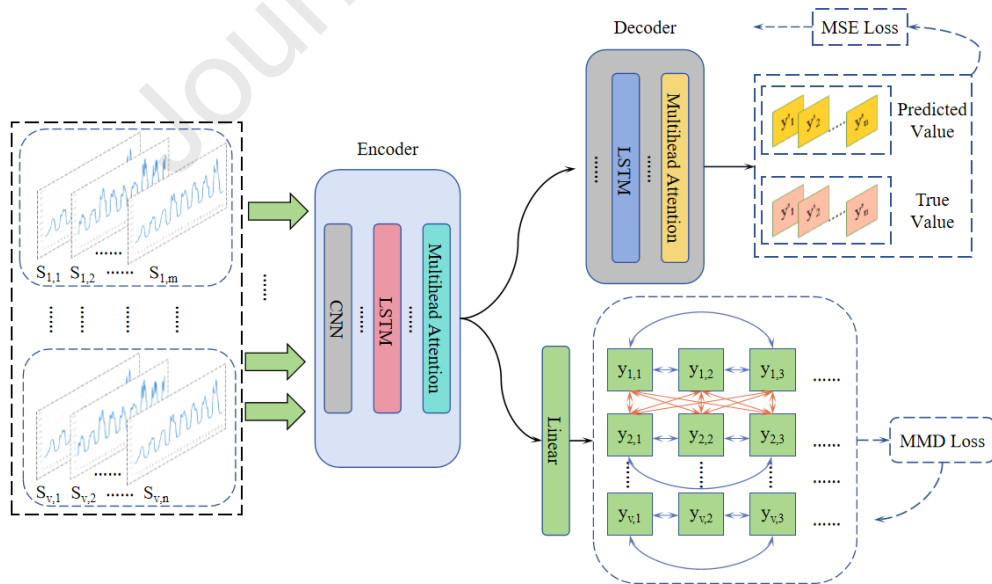


Fig.4. Multi-source generalization model architecture

In this study, MSE will be used to calculate the error loss (MSE_{pred}) between the predicted values and the real data, and MMD will be used to calculate the distance loss, the formula for MMD is as follows.

$$d_{mmd}(x_a, x_b) = \frac{1}{n_a} \sum_{i=1}^{n_a} \sum_{j=1}^{n_a} k(x_{ai}, x_{aj}) + \frac{1}{n_b} \sum_{i=1}^{n_b} \sum_{j=1}^{n_b} k(x_{bi}, x_{bj}) - \frac{2}{n_a n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} k(x_{ai}, x_{bj}) \quad (16)$$

Where x_a and x_b are two different sets of data and k is the kernel function, in this study we will use Radial Basis Function (RBF) as k . The loss function of the whole multi-source generalization model is as follows.

$$l_{total} = \frac{1}{M} \left(\sum_{i=1}^M MSE_{pred,i} \right) + \lambda \frac{1}{N} \left(\sum_{j=1}^N d_{mmd,j} \right) \quad (17)$$

Where l_{total} is the total loss, M denotes the total number of segments in the whole source domain after the division by TDC, and N is the number of do not repeat combinations within and between domains. $\lambda = \frac{2}{1+e^{-10 \frac{p}{iteration}}} - 1$, $0 \leq p \leq iteration$ [37]. To contain instability in the early stages of training,

λ was gradually changed from 0 to 1. The optimizer Adam is utilized during the model training process, setting the maximum number of iterations, loss value thresholds, and a method to adjust the learning rate by monitoring the variation of losses to prevent model overfitting.

3. Case Study

3.1 Data Source

Real datasets will be used for testing to validate the ability of the multi-source domain generalization model used in this study to perform the energy consumption prediction task in both the source and target domains. The datasets are derived from energy consumption data of office buildings from a school in Dalian, Liaoning Province, China, and the building information is presented in Table 1. Since the energy consumption of air-conditioning systems accounts for a large proportion of total building energy consumption, the form of air-conditioning was also used as a screening condition for buildings of the same type. The air-conditioning system in each building in the selected dataset is a split air-conditioner, and the data collection interval is 1 hour. Although some of the buildings were collected in different years, they were in the same months, January through July, which does not affect the prediction of energy consumption trends. Among these buildings, Buildings 1-4, 6 and 9 are multidisciplinary research office buildings containing laboratories. Compared to the other three office buildings, the main energy consumption includes laboratory equipment in addition to lighting outlets, air-conditioning systems, office equipment, and power equipment energy consumption. These buildings are used for office from 9 a.m. to 5 p.m., but energy consumption is generated outside of office hours due to student research and laboratory use, which increases the volatility of daily energy consumption data and makes it more difficult to predict energy consumption trends. This study also considers whether the generalization model can be used directly for energy consumption prediction for a similar building when there is little or no training data for that building (such as a newly constructed building). For this purpose, energy consumption prediction experiments were designed for two separate buildings, that is, Buildings 8 and 9. These buildings are not involved in the source domain data screening process and is directly classified to the target domain. The effectiveness of multi-source generalization models in the task of energy consumption prediction for such buildings is explored on this two building test set.

Table 1

Office Building Information

Name	Storey	Building area (m ²)	Purpose	Air conditioning form	Acquisition time
Building1	10	25000	Office and research	Split air conditioner	2022.1-2022.7
Building 2	7	11000	Office and research		2021.1-2021.7
Building 3	5	9533	Office and research		2021.1-2021.7
Building 4	6	35998	Office and research		2021.1-2021.7
Building 5	5	5188	Office		2022.1-2022.7
Building 6	1	547	Office and research		2021.1-2021.7
Building 7	7	8514	Office		2023.1-2023.7
Building 8	7	17540	Office and teach		2021.1-2021.7
Building 9	5	10196	Office and research		2023.1-2023.7

In the real dataset, several consecutive segments of missing values are generated for reasons such as power outages, maintenance and sensor failure, which are difficult to be filled in simply, so these missing values are processed for delete. There are also differences in the temporal distribution due to the different years of data sources and time periods of missing data for different building energy consumption datasets. However, even for buildings of the same type in the same area, the change in energy consumption is mainly affected by their own characteristics and the way they are used, so this phenomenon has limited impact on the validation of the generalization model. These factors may also increase the complexity of model training and prediction. Fig.5 described the percentage of energy consumption values in different intervals for each building after the missing value treatment. It can be observed that the distribution of energy use and fluctuation intervals of each building in the same type also have some differences.

In this study, the first 4200 sets of data from all buildings were used as a dataset, of which 80% was used to train the model and 20% was used to test the prediction of the model. When dividing data sets, we should also try to ensure that the starting time points of the test sets in different buildings are close to each other. Without considering the year, the maximum interval between the starting points of different building energy consumption data test sets is about 2 day, accounting for a small proportion of the test set. So the difference in the starting point of the test set does not have a large impact on the validation of the generalization model in this research. And the energy consumption data of the previous one and two hours are used as inputs to the model to predict the energy consumption of the next hour. Finally the results are compared on a test set of all buildings.

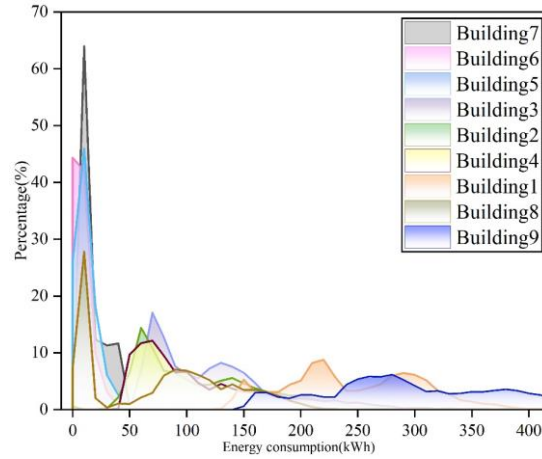


Fig.5. Energy consumption ratio curve

3.2 Experimental Design

Fig.6 illustrates the objective of this study, which aims to predict energy consumption for office buildings one hour in advance by filtering the source domain data in the dataset and training a generalization model applicable for energy consumption prediction in both the source and target domains. To demonstrate the predictive power of the generalization model, various comparison tests will be established in this study.

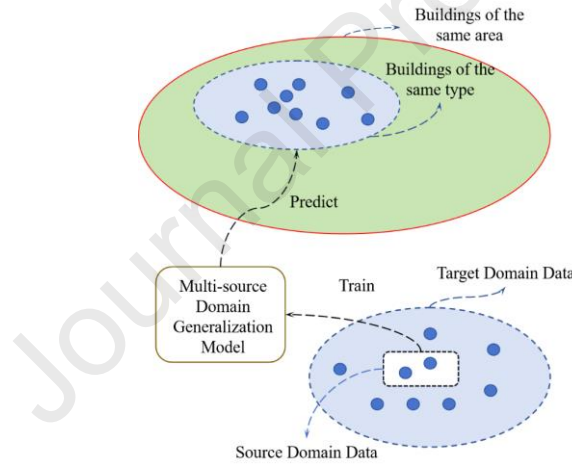


Fig.6. Goals of a multi-source domain generalization model

Initially, the prediction results were first compared on the source domain test set with prediction models trained using a single architecture including CNNLSTM, Bi-directional Long Short-Term Memory (BiLSTM), two LSTMs (2LSTM), LSTM, and CNN-AttLSTM over the source domain to see if the TDC-CNN-AttLSTM outperforms the other comparison models or achieves an acceptable range of results.

In addition, a comparison experiment with the prediction of a CNN-AttLSTM model trained using a single source domain on a test set in the target domain was designed. This aims to assess whether the predictive model on the single source domain can be directly applied to the target domain, demonstrating the generalization predictive ability of the multi-source generalization model. Finally, experiments were also conducted on two separate building that did not participate in the source data screening to explore the effectiveness of the multi-source generalization model in predicting energy consumption on the same type of building with an unknown training set. The overall technical route is shown in Fig.7.

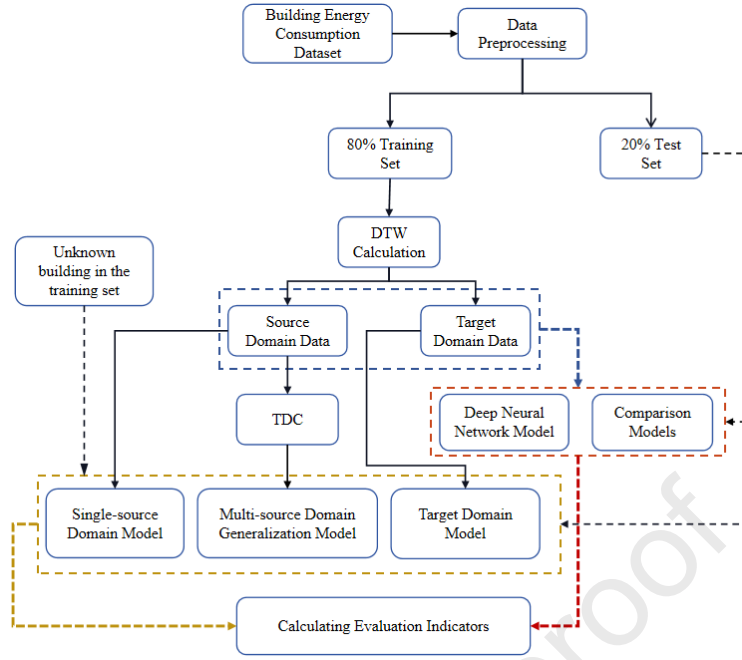


Fig.7. Technological line

3.3 Data Preprocessing

In addition to the missing value processing as described in 3.1 and dataset division, the data need to be normalized or standardized to improve the speed and accuracy of the model training. This study will carry out the MinMax normalization operation on the input parameters, which is calculated by the following formula.

$$x_{std} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (18)$$

$$x_{std2} = x_{std} (x_{max} - x_{min}) + x_{min} \quad (19)$$

Where x_{std} means the data after normalization, x_{max} and x_{min} means the maximum and minimum values in the parameter respectively, and finally the inverse normalization calculation is performed to get the final calculation result x_{std2} .

After normalization calculations, the source domain buildings need to be filtered by DTW calculations on the training set of all buildings. As shown in the Fig.8, it is the result after performing the DTW calculation and normalizing.

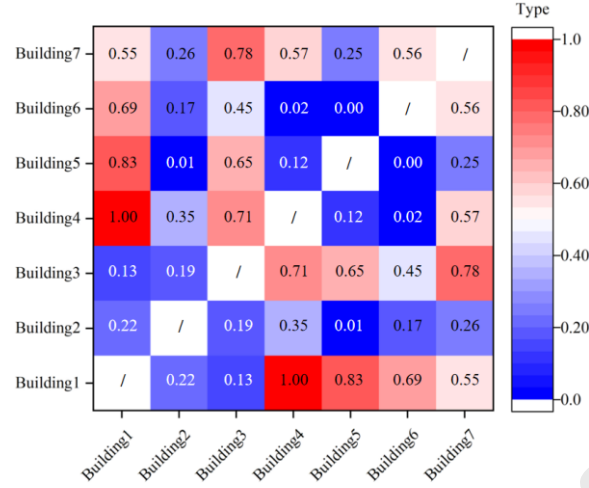


Fig.8. DTW calculation results

The top three buildings with the smallest overall distance from others were selected by calculating the mean value of each column. And two source domain buildings were further screened based on variance to prevent the generalization model from learning information that could negatively affect the prediction of energy consumption trends in other buildings during the training process. The calculation results are presented in Table 2. Ultimately, building 2 and building 6 were chosen as the source domain buildings. Building 2, which had the minimum mean value calculation, was partitioned into 3 parts, while building 6 was partitioned into 2 parts using TDC.

Table 2

Calculation of mean and variance of DTW for each building against other buildings

	Building 1	Building 2	Building 3	Building 4	Building 5	Building 6	Building 7
Average	0.57	0.20	0.49	0.46	0.31	0.32	0.50
Variance	0.12	0.01	0.08	0.14	0.12	0.09	0.04

3.4 Hyperparameter Optimization

Hyperparameter optimization is very important for black-box models, and determination of each hyperparameter is another very complex task for model training[56]. Bayesian optimization will use the performance effects of previously searched parameters to infer the next parameter search as a way to improve search efficiency[57]. Tree-structured Parzen Estimator (TPE) Bayesian optimization uses a Gaussian mixture model to optimize the hyperparameters. In Bayesian theory, $p(x | y)$ is the probability of the hyperparameter x when the loss is y [58]. That parameter is defined in a different way by TPE.

$$p(x|y) = \begin{cases} l(x) & y < y^* \\ g(x) & y \geq y^* \end{cases} \quad (20)$$

Where $l(x)$ is the probability density observed under x when the loss of $f(x)$ is less than y^* and $g(x)$ is the probability density formed by the remaining observations.

In the original TPE Bayesian optimization text, the original EI formulation is simplified and the optimized results are shown below.

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} y p(y) dy}{\gamma l(x) + (1 - \gamma) g(x)} \propto \left(\gamma + \left(\frac{g(x)}{l(x)} \right) (1 - \gamma) \right)^{-1} \quad (21)$$

As can be seen in the final equation, the TPE will use $g(x)/l(x)$ to evaluate all candidate hyperparameter combinations. When the number of iterations is completed or the predefined abort condition is reached, the candidate hyperparameter combination x^* with the largest EI value will be returned to finalize the hyperparameter optimization.

In this study, the evaluation was done by maximizing the correlation coefficient (R^2) of the test set and the upper limit of the search number is set to 20. The Table 3 shows the range of hyperparameter optimization for each model in this study and the Table 4 shows the final optimization results for each building. Where the number 1-7 corresponds to Building1-7 respectively, the multi-source generalization model is used to predict energy consumption for all buildings so there is only one set of hyperparameters. Since the results of hyperparameter optimization can vary greatly with different datasets or different computing devices, they need to be considered on a practical basis when using hyperparameter optimization.

Table 3

Hyperparameter optimization range

Model	Parameters	Optimization Scope	Partition
TDC-ATT-CNNLSTM	Encoder LSTM size	[200,400]	20
	Decoder LSTM size	[200,400]	20
	Learning rate	[0.0001,0.002]	0.0001
ATT-CNNLSTM	Encoder LSTM size	[200,400]	20
	Decoder LSTM size	[200,400]	20
	Learning rate	[0.0001,0.002]	0.0001
CNNLSTM	LSTM size	[200,400]	20
	Learning rate	[0.0001,0.002]	0.0001
BiLSTM	LSTM size	[300,400]	20
	Learning rate	[0.0001,0.002]	0.0001
2LSTM	First LSTM size	[300,400]	20
	Second LSTM size	[300,400]	20
	Learning rate	[0.0001,0.002]	0.0001
LSTM	LSTM size	[300,400]	20
	Learning rate	[0.0001,0.002]	0.0001

Table 4

Hyperparameter optimization results

Model	Parameters	1	2	3	4	5	6	7
TDC-	Encoder LSTM size				360			
ATT-	Decoder LSTM size				360			
CNNLSTM	Learning rate ($\times 10^{-2}$)				0.19			
ATT-	Encoder LSTM size	400	200	280	200	280	220	260
CNNLSTM	Decoder LSTM size	260	400	260	360	380	280	240
M	Learning rate ($\times 10^{-2}$)	0.14	0.18	0.12	0.18	0.16	0.14	0.14
CNNLSTM	LSTM size	360	380	320	400	400	360	400
M	Learning rate ($\times 10^{-2}$)	0.17	0.04	0.07	0.10	0.09	0.06	0.10
BiLSTM	LSTM size	320	360	300	300	300	360	400

	Learning rate (*10 ⁻²)	0.19	0.16	0.13	0.07	0.18	0.06	0.20
2LSTM	First LSTM size	360	300	300	300	400	340	380
	Second LSTM size	360	300	320	320	360	340	380
	Learning rate (*10 ⁻²)	0.05	0.18	0.05	0.11	0.07	0.13	0.04
LSTM	LSTM size	320	360	300	360	320	380	340
	Learning rate (*10 ⁻²)	0.15	0.20	0.18	0.15	0.09	0.10	0.10

3.5 Evaluation Indicators

In this study, mean absolute error(MAE), Root mean square error coefficient of variation(CV-RMSE) and R^2 will be used to evaluate the predictive performance of the model[59-61]. The calculation formula is as follows.

$$MAE = \frac{\left(\sum_{i=1}^{i=n} |f(x_i) - y_i| \right)}{n} \quad (22)$$

$$CV - RMSE = \sqrt{\frac{\sum_{i=1}^n (f(x_i) - y_i)^2}{n}} / \frac{\sum_{i=1}^n y_i}{n} \quad (23)$$

$$R^2 = 1 - \frac{\left(\sum_{i=1}^{i=n} (f(x_i) - y_i)^2 \right)}{\left(\sum_{i=1}^{i=n} (y_i - \bar{y})^2 \right)} \quad (24)$$

Where $f(x)$ represents the data predicted by the model, y represents the real data, and n is the number of data. The closer MAE and CV-RMSE are to 0, and the closer R^2 is to 1 indicates that the closer the trend of the overall prediction results and the real values are, the better the prediction effect of the model will be.

4. Results and Discussion

4.1 Discussion of Source Domain Prediction Results

Fig.9 and Fig.10 display the prediction results on the source domain test set. Observing the prediction results at different time periods reveals that all the prediction models on the two buildings in the source domain roughly capture the energy consumption trends. However, near some energy use peaks and troughs all the prediction models deviate from the true values to different degrees.

Upon assessing the different error ranges from the Fig.11. It is observed that the prediction error distribution of TDC-CNN-AttLSTM is primarily within 10kWh, whereas the prediction errors of other models exhibit a high percentage in other ranges, particularly evident in the prediction results of source building 2. Overall, it is evident that the multi-source domain generalization model can more accurately predict energy consumption trends on the test set of the two source domain buildings.

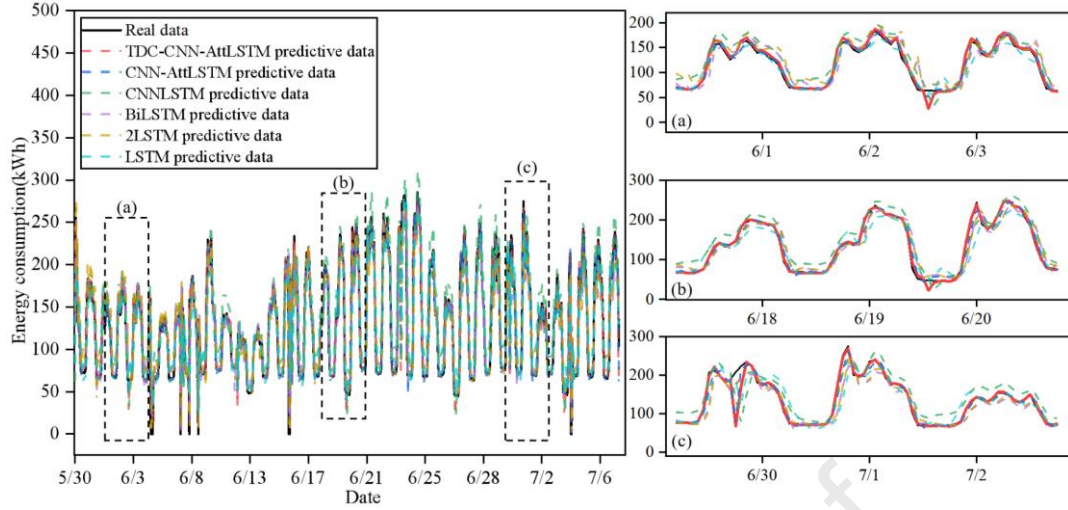


Fig.9. Building 2 energy consumption prediction results

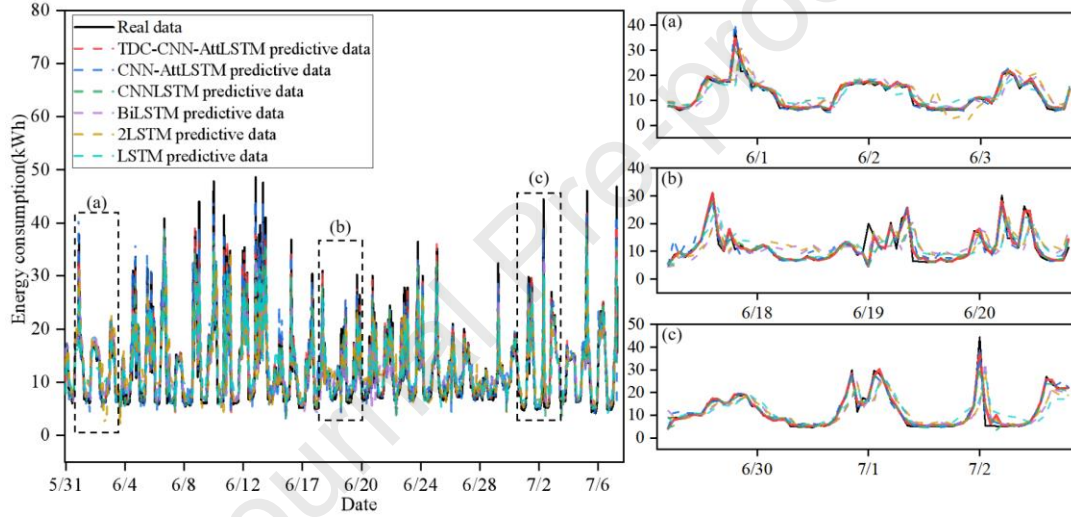


Fig.10. Building 6 energy consumption prediction results

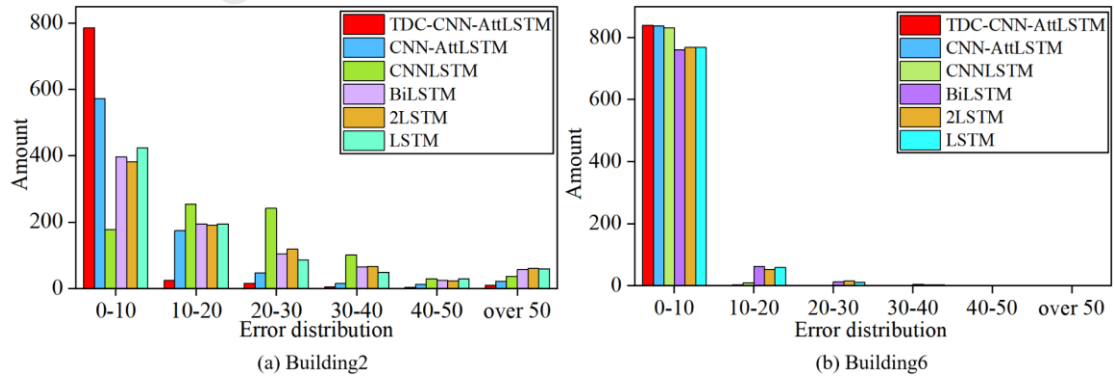


Fig.11. Distribution of errors in source domain prediction results

Table 5 presents the results of various models for evaluating the predicted and true values on the source domain test set. The computational findings reveal that both TDC-CNN-AttLSTM and CNN-AttLSTM outperform the other models on both buildings, achieving an R^2 higher than 0.9. The R^2 improvement on the two source domain building test set compared to the best results in the other comparison models ranges from 3.1% to 21.9%. Additionally, the multi-source domain

generalization model, combined with TDC, exhibits slightly superior results compared to those of the deep neural network, both of which improve R^2 by more than 4%. This underscores the acquisition of more temporal invariant knowledge by the model during the generalization model training process, thereby contributing to its enhanced performance in predicting results.

Table 5

Calculation of source domain test set evaluation metrics

Name	Evaluation metrics	TDC-CNN-AttLSTM	CNN-AttLSTM	CNNLSTM	BiLSTM	2LSTM	LSTM
Building2	MAE	4.747	9.955	21.979	18.210	18.575	17.582
	CV-RMSE	0.090	0.135	0.217	0.225	0.228	0.226
	R^2	0.964	0.919	0.791	0.776	0.768	0.772
Building6	MAE	1.021	1.332	1.299	3.863	3.817	3.738
	CV-RMSE	0.145	0.156	0.187	0.508	0.501	0.488
	R^2	0.945	0.936	0.908	0.322	0.340	0.374

4.2 Discussion of Target Domain Prediction Results

The multi-source domain generalization model is tested on the target domain test set to demonstrate the generalization ability of the model. The prediction results on the target domain buildings 1, 3-5, and 7 are depicted in Figs.12-16. Upon conducting global and local zoom in observations, it is evident that the overall prediction effect is similar to the results on the source domain, with a small number of larger errors generated in the prediction results of all models. This may be attributed to irregular energy use by the students, as well as data incoherence due to power outages, which can impact the prediction results. However, this effect was quickly eliminated, demonstrating the adaptive ability of the prediction models.

Overall, the predictions of the multi-source domain generalization model exhibit closer alignment with the true values for different buildings. In terms of error distribution statistics from the Fig.17, TDC-CNN-AttLSTM outperforms other prediction models in most target domain buildings, with errors primarily within 10kWh. Although a significant number of errors are distributed within 20kWh for building 4, the calculations of the other models contribute to a larger portion of higher error intervals. Furthermore, observation of the partial zoomed in plots reveals that the generalization model offers a more accurate description of the energy consumption trend compared to the other models.

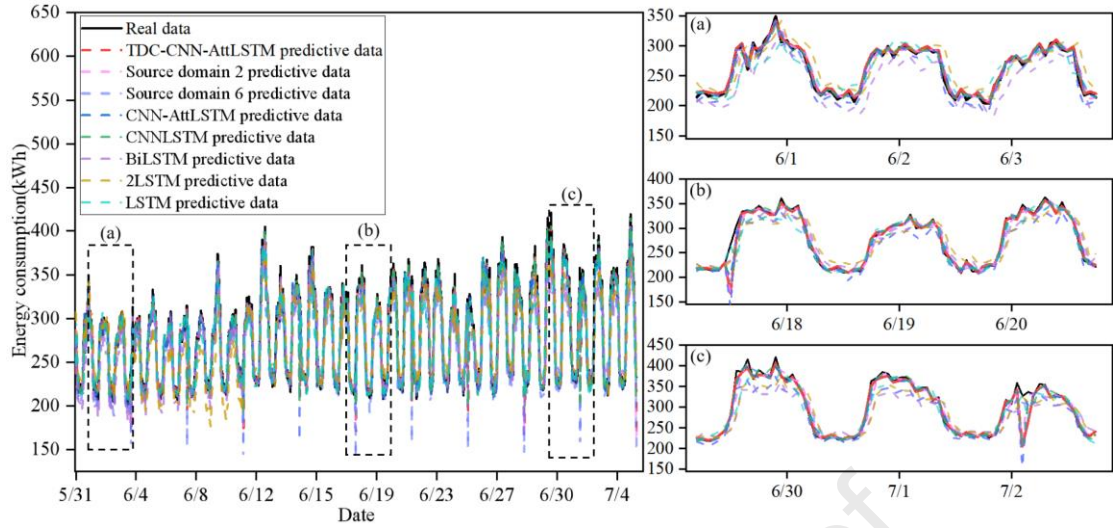


Fig.12. Building 1 energy consumption prediction results

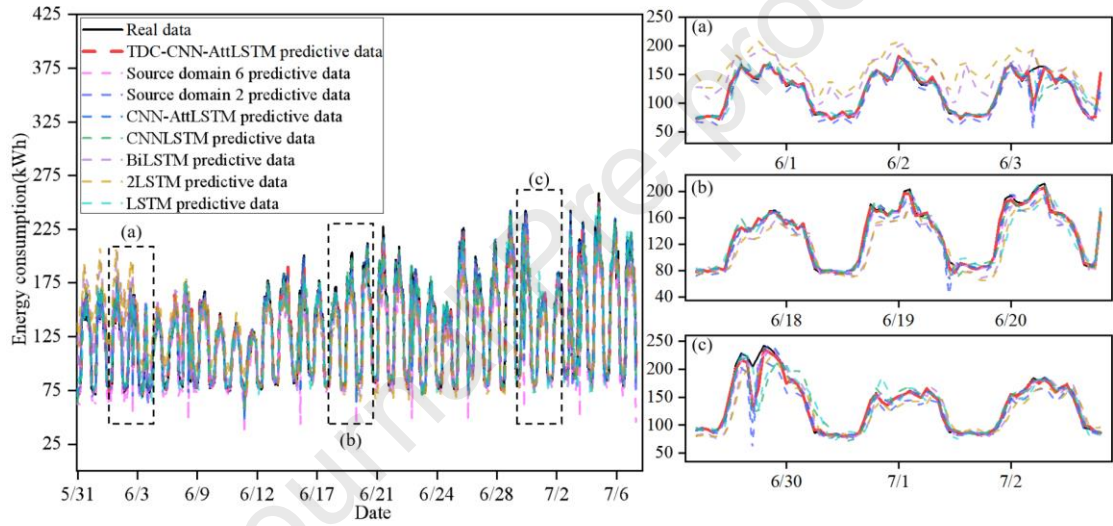


Fig.13. Building 3 energy consumption prediction result

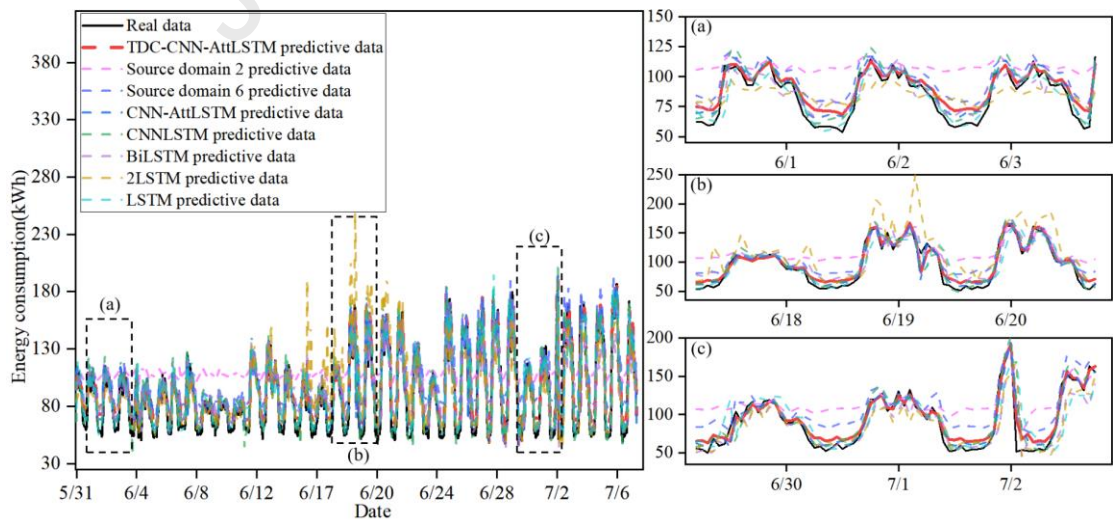


Fig.14. Building 4 energy consumption prediction results

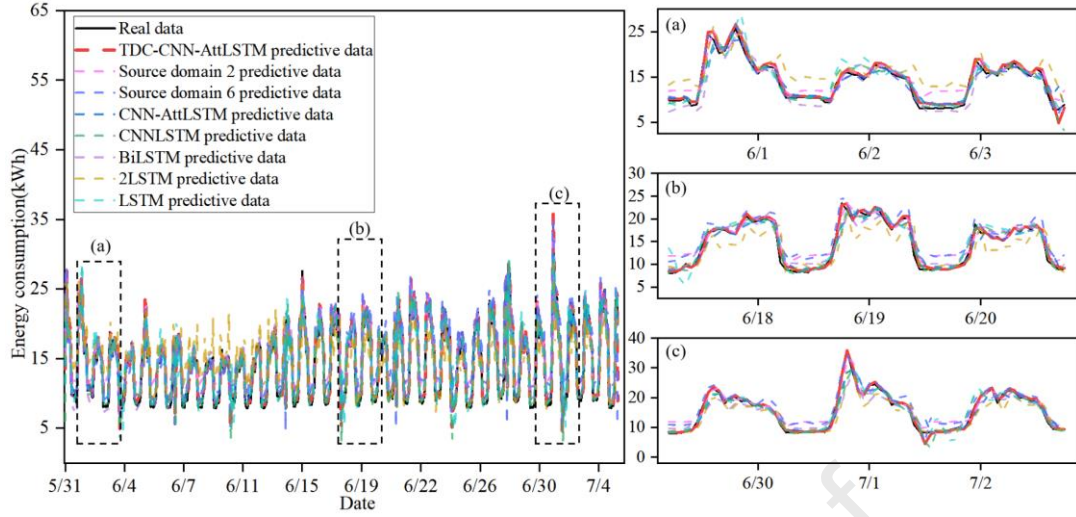


Fig.15. Building 5 energy consumption prediction results

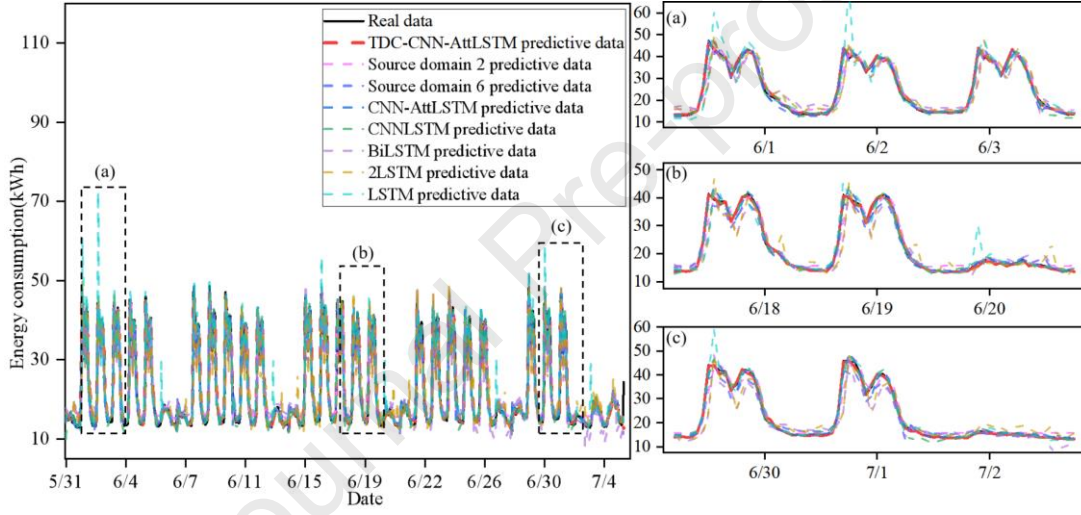


Fig.16. Building 7 energy consumption prediction results

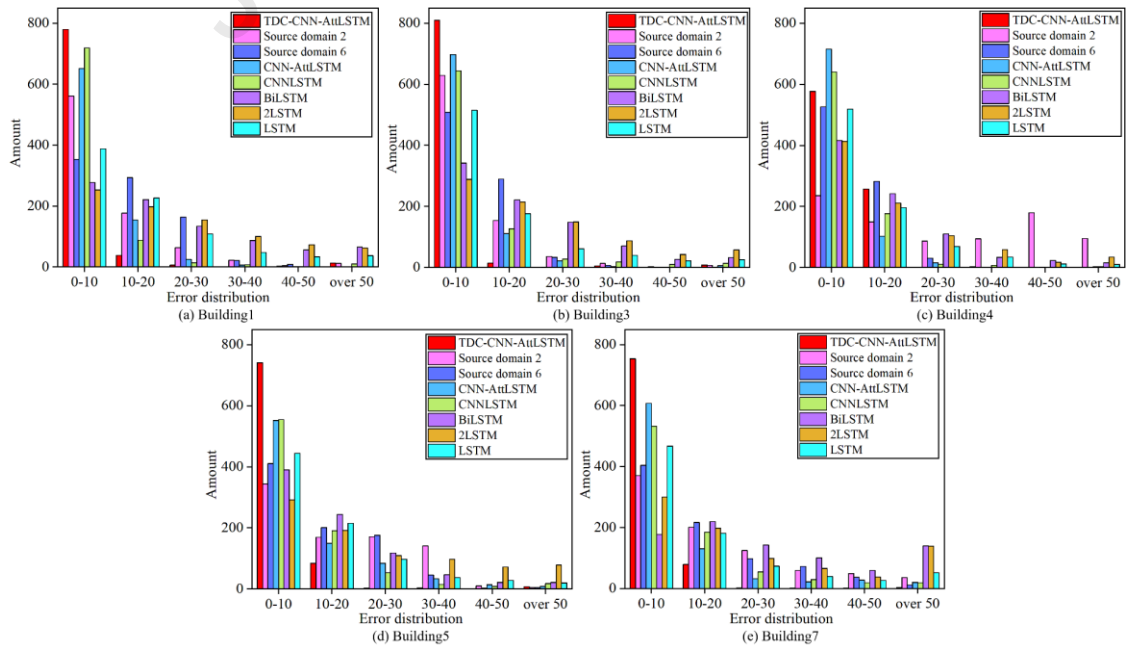


Fig.17. Distribution of errors in target domain prediction results

Table 6 shows the results of the evaluation metrics computed on the test set in the target domain. It can be found that the TDC-CNN-AttLSTM model basically has better prediction results on each building, better than the results of the direct prediction using only the trained single-source domain prediction model(Source domain 2, Source domain 6) and better or close to the prediction results of the model trained with the target domain(CNN-AttLSTM). It is shown that multiple source domain training and the TDC algorithm largely enhance the extraction of domain invariant features and improve the predictive power of the generalization model, which cannot be obtained by using the source domain predictive model directly on the target domain. And the comparison with other model prediction results found that the generalization model prediction results are all basically superior. Although the R^2 on building 1 is slightly lower than the results of CNNLSTM, they all reach more than 0.9, which is an acceptable range.

Table 6**Calculation of evaluation metrics on the source domain test set**

Name	Evaluation metrics	TDC-CNN-AttLSTM	Source domain 2	Source domain 6	CNN-AttLSTM	CNNLSTM	BiLSTM	2LSTM	LSTM
Building1	MAE	5.484	10.093	14.694	6.949	6.411	20.717	22.398	15.815
	CV-RMSE	0.047	0.057	0.080	0.034	0.046	0.097	0.101	0.079
	R^2	0.936	0.905	0.816	0.966	0.938	0.729	0.705	0.820
Building3	MAE	3.574	7.648	10.112	5.960	8.292	16.702	19.715	12.111
	CV-RMSE	0.075	0.091	0.121	0.084	0.108	0.167	0.197	0.142
	R^2	0.951	0.927	0.871	0.938	0.898	0.756	0.660	0.824
Building4	MAE	6.880	26.118	14.007	6.311	7.385	13.421	14.968	10.886
	CV-RMSE	0.097	0.347	0.188	0.099	0.107	0.199	0.236	0.170
	R^2	0.931	0.119	0.742	0.929	0.916	0.710	0.594	0.790
Building5	MAE	0.587	1.677	1.488	1.061	1.000	1.465	2.222	1.328
	CV-RMSE	0.080	0.150	0.142	0.105	0.117	0.141	0.206	0.134
	R^2	0.950	0.825	0.843	0.914	0.893	0.846	0.668	0.860
Building7	MAE	0.434	1.810	1.435	0.966	1.103	3.190	2.868	1.729
	CV-RMSE	0.034	0.124	0.090	0.072	0.073	0.212	0.212	0.151
	R^2	0.995	0.931	0.964	0.977	0.976	0.800	0.798	0.898

Fig.18 shows the statistical results of the evaluation indexes of various prediction models on the source and target domains. Through the mapping of different axes, it can be found that the prediction results of the multi-source domain generalization model on both source and target domains is better than others in most cases, and the distribution is more dense. It shows that the model can have good generalization prediction ability for both source and target domains, and the prediction results are more stable.

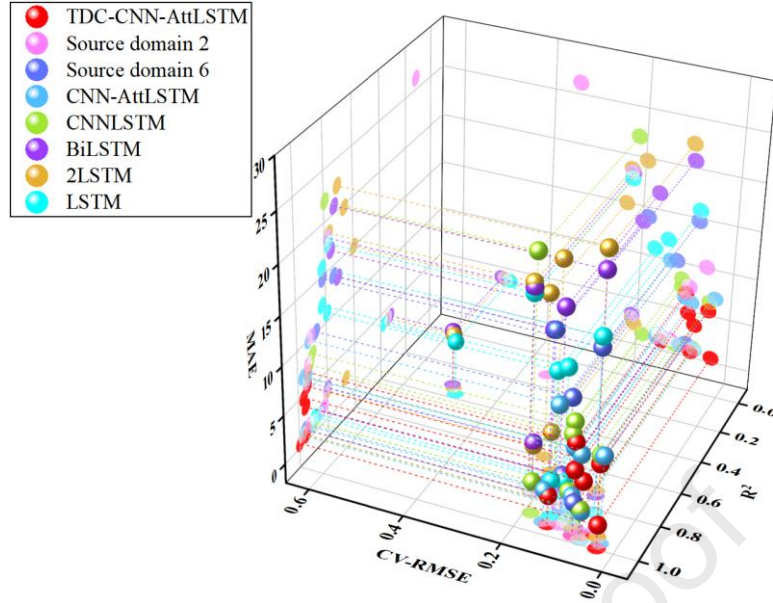


Fig.18. Distribution of computed results of evaluation metrics on the test set

4.3 Discussion of results for target domains with unknown training sets

In the following, we will explore whether the multi-source generalization model can be directly used for the same type of building energy consumption prediction task in the same area where the training set is unknown. We will test its effectiveness on two buildings that do not participate in source domain screening. Since the above experiments have already demonstrated the prediction effectiveness of CNN-AttLSTM on multiple buildings in both the source and target domains, only the deep neural network trained in the source domain is set up here to be used directly for comparison experiments on this building test set. As shown in Fig. 19 and Fig. 20, these are the predicted energy consumption results for Buildings 8 and 9. Different error lines have been chosen for Buildings 8 and 9 due to their different data distributions. Building 8 will use a $\pm 10\%$ error line, while building 9 uses a $\pm 10\text{kWh}$ error line. Table 7 shows the percentage of predicted values outside the error line for the different prediction methods, with the generalization model having a significantly lower percentage than the other two compared methods. In the generalization model prediction results, the points outside the error line are largely distributed around the daily extremes. It shows that while the generalization model is able to predict the general trend of energy consumption, its effectiveness in predicting near the extreme points needs to be improved. As shown in Table 8, it is the calculation result of the evaluation index. It can be found that the evaluation metrics of the multi-source generalization model still outperform the results of the direct prediction of the deep neural network trained using only the source domain, with the improvement 5.6% - 26.6% in R^2 . Therefore, the multi-source generalization model also has a certain adaptability to some of the same type of building energy consumption prediction tasks where the training set is unknown.

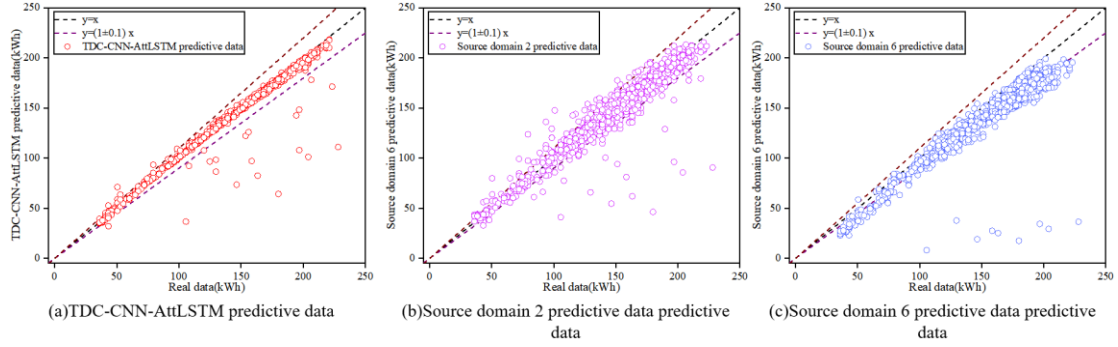


Fig.19. Building 8 energy consumption prediction results

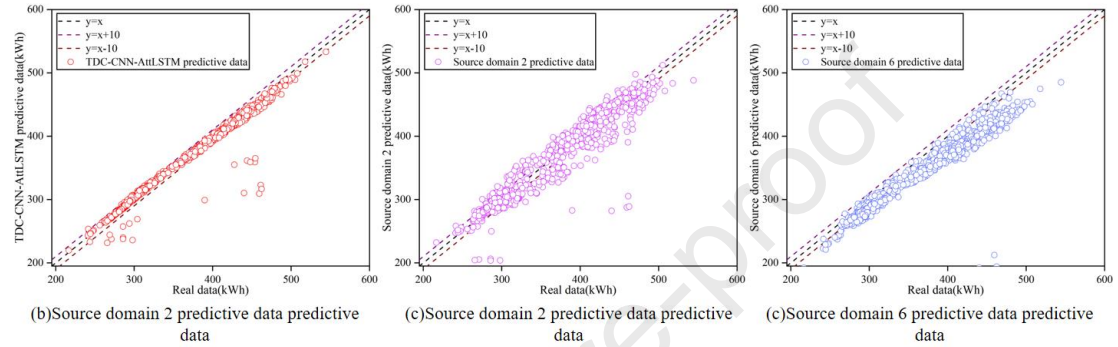


Fig.20. Building 9 energy consumption prediction results

Table 7

Percentage of forecast error outside the error line

Name	TDC-CNN-AttLSTM	Source domain 2	Source domain 6
Building8	3.33%	14.52%	41.43%
Building9	11.55%	46.31%	78.57%

Table 8

Calculation of evaluation metrics on the building 8 test set

Name	Evaluation metrics	TDC-CNN-AttLSTM	Source domain 2	Source domain 6
Building 8	MAE	3.630	7.696	14.026
	CV-RMSE	0.073	0.101	0.149
	R ²	0.946	0.895	0.772
Building 9	MAE	5.816	13.226	24.210
	CV-RMSE	0.039	0.057	0.093
	R ²	0.958	0.907	0.757

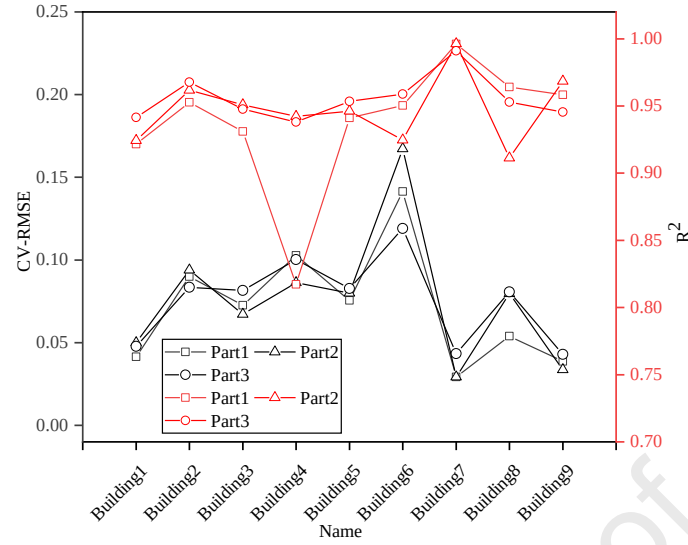


Fig.21. Analysis of local prediction results of generalization model

In order to verify the stability of the generalization model in different time periods, the test set of nine buildings in the target and source domains is divided evenly into three parts in this paper. The results of the evaluation metrics CV-RMSE and R^2 for the prediction results in different parts are shown in Fig.21. Most of the results show that the two evaluation metrics are very close to each other on the different parts of the calculation, except for the third part of Building 4 (Part1) where R^2 is lower and on Building 6 (Part2) where CV-RMSE is more dispersed. Part of the reason for this phenomenon may be due to data completeness. With Fig. 14(a), we find that the performance of the generalization model decreases slightly near the extreme points, which also leads to a lower R^2 . By zooming in locally in Fig.10(b), we find that the energy consumption of building 6 has a small range and its fluctuations are more sensitive to large-scale energy use cases such as experiments. The generalization model cannot deal well with the prediction of these points near extreme. Therefore, there is still much room for improvement in the training and validation of generalization models.

5. Conclusion

In pursuit of a generalization model for predicting energy consumption across multiple buildings of the same type in both the source and target domains without accessing target domain data during model training, a multi-source domain generalization model TDC-CNN-AttLSTM was developed in this study. This model was trained on a real dataset using a combination of deep neural network and transfer learning techniques. In the one hour ahead office building energy prediction task, the DTW is used to obtain source domain buildings and train a generalization model, which is tested and evaluated metrics are computed on a test set of two source domain buildings, five target domain buildings, and two buildings with an unknown training set. The conclusions reached are as follows.

(1) The results of CNN-AttLSTM on the source domain test set are better than the other compared models, indicating the effectiveness of the deep neural network for the energy consumption prediction task on the datasets used in this study. And the prediction results of TDC-CNN-AttLSTM are all better than CNN-AttLSTM with the lowest R^2 of 0.94 on the source domain test set. It indicates that TDC decomposition and multi-source domain generalization training enhance the prediction ability of the model.

(2) The prediction results of the TDC-CNN-AttLSTM model on most of the buildings on the target domain are also better or close to the prediction results of the CNN-AttLSTM, and although the evaluation indexes calculated on some buildings are slightly lower compared with a few models, the R^2 is also above 0.93, which is an acceptable range. It shows that the multi-source domain generalization model has a basic ability to generalize on the dataset of this study. It can be used for energy consumption prediction task for multiple buildings of the same type in the same area.

(3) By comparing the TDC-CNN-AttLSTM model on the target domain test set with the CNN-AttLSTM model trained on a single source domain, it is found that the prediction results of the multi-source domain generalization model are better. The results show that energy consumption prediction models trained with selected single-source domain data cannot be directly used in the target domain on this dataset, which proves the effectiveness of multi-source generalization models.

(4) Finally, using the trained generalization model directly on the task of energy consumption prediction for a target building of the same type with an unknown training set. It is found that the multi-source generalization model can still predict it well with an R^2 has improved 5.6% - 26.6%. It shows that the generalization model still has some prediction ability for the energy consumption prediction task for some buildings with small datasets or unknown training sets.

However, the source domain selection process and the overall architecture of the model still need to be optimized. Furthermore, the performance of the generalization model on richer and more complete real datasets still needs to be tested. Training generalization model on real datasets from other building groups still needs to be explored and more factors may need to be considered. For example, in commercial buildings, the way of energy using equipment operation, rationalization of the source domain selection process, applicability of the network architecture, and the richness of the dataset need to be taken into account.

Acknowledgment

The authors are grateful for the financial support of the National Natural Science Foundation of China (Grant No. 52078096 and 51806029), 2023 Dalian University of Technology - Cardiff University Cooperation and Exchange Foundation Project, and 2023 International Exchange Foundation Project of "Co-Creation of Excellence Program" from Dalian University of Technology (Grant: DUTIO-ZG-202307), the Key Project of DUT for International Students Studying and Researching in China: Innovation and Practice of Talent Cultivation Model in the Field of Smart Buildings for the "Belt and Road" Initiative (Grant No. 1103-82120001) and the Natural Science Foundation Joint Fund of Liaoning Province (Grant No. 2023-MSBA-023).

References

- [1] Li Q, Meng Q, Cai J, et al. Applying support vector machine to predict hourly cooling load in the building[J]. *Applied Energy*, 2009, 86(10): 2249-2256.
- [2] Zhao H, Magoules F. A review on the prediction of building energy consumption[J]. *Renewable and Sustainable Energy Reviews*, 2012, 16(6): 3586-3592.
- [3] Foucquier A, Robert S, Suard F, et al. State of the art in building modelling and energy performances prediction: A review[J]. *Renewable and Sustainable Energy Reviews*, 2013, 23: 272-288.
- [4] Zhang Y, Teoh B K, Zhang L. Exploring driving force factors of building energy use and GHG emission using a spatio-temporal regression method[J]. *Energy*, 2023, 269: 126747.
- [5] Amasyali K, El-Gohary N M. A review of data-driven building energy consumption prediction studies[J]. *Renewable and Sustainable Energy Reviews*, 2018, 81: 1192-1205.
- [6] Wei Y, Zhang X, Shi Y, et al. A review of data-driven approaches for prediction and classification of building energy consumption[J]. *Renewable and Sustainable Energy Reviews*, 2018, 82: 1027-1047.
- [7] Zhang Y, Bai X, Mills F P, et al. Rethinking the role of occupant behavior in building energy performance: A review[J]. *Energy and Buildings*, 2018, 172: 279-294.
- [8] S. J P, Q. Y. A Survey on Transfer Learning[J]. *Ieee Transactions On Knowledge and Data Engineering*, 2010, 22(10): 1345-1359.
- [9] Ahmad A S, Hassan M Y, Abdullah M P, et al. A review on applications of ANN and SVM for building electrical energy consumption forecasting[J]. *Renewable and Sustainable Energy Reviews*, 2014, 33: 102-109.
- [10] Ahmad M W, Mourshed M, Rezgui Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption[J]. *Energy and Buildings*, 2017, 147: 77-89.
- [11] Aydinalp M, Ismet Ugursal V, Fung A S. Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks[J]. *Applied Energy*, 2002, 71(2): 87-110.
- [12] Rahman A, Srikumar V, Smith A D. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks[J]. *Applied Energy*, 2018, 212: 372-385.
- [13] Somu N, M R G R, Ramamritham K. A hybrid model for building energy consumption forecasting using long short term memory networks[J]. *Applied Energy*, 2020, 261: 114131.
- [14] Fan C, Wang J, Gang W, et al. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions[J]. *Applied Energy*, 2019, 236: 700-710.
- [15] Kim T, Cho S. Predicting residential energy consumption using CNN-LSTM neural networks[J]. *Energy*, 2019, 182: 72-81.
- [16] Yang Y, Hong W, Li S. Deep ensemble learning based probabilistic load forecasting in smart grids[J]. *Energy*, 2019, 189: 116324.
- [17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in Neural Information Processing Systems*, 2017, 30.
- [18] Li Y, Tong Z, Tong S, et al. A data-driven interval forecasting model for building energy prediction using attention-based LSTM and fuzzy information granulation[J]. *Sustainable*

- Cities and Society, 2022, 76: 103481.
- [19] Ding Z, Chen W, Hu T, et al. Evolutionary double attention-based long short-term memory model for building energy prediction: Case study of a green building[J]. *Applied Energy*, 2021, 288: 116660.
 - [20] Gao Y, Ruan Y. Interpretable deep learning model for building energy consumption prediction based on attention mechanism[J]. *Energy and Buildings*, 2021, 252: 111379.
 - [21] Chen Y, Tong Z, Zheng Y, et al. Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings[J]. *Journal of Cleaner Production*, 2020, 254: 119866.
 - [22] Ma J, Cheng J C P, Lin C, et al. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques[J]. *Atmospheric Environment*, 2019, 214: 116885.
 - [23] Ma J, Li Z, Cheng J C P, et al. Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network[J]. *Science of the Total Environment*, 2020, 705: 135771.
 - [24] Jason Y, Jeff C, Yoshua B, et al. How transferable are features in deep neural networks?[J]. *Advances in Neural Information Processing Systems*, 2017, 27.
 - [25] Bayer M, Frey T, Reuter C. Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence[J]. *Computers & Security*, 2023, 134: 103430.
 - [26] Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks: International conference on machine learning, 2017[C]. PMLR.
 - [27] Oyewole I, Chehade A, Kim Y. A controllable deep transfer learning network with multiple domain adaptation for battery state-of-charge estimation[J]. *Applied Energy*, 2022, 312: 118726.
 - [28] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. *The Journal of Machine Learning Research*, 2016, 17(1): 2030-2096.
 - [29] Ouyang X, Yang Y, Zhang Y, et al. Domain adversarial graph neural network with cross-city graph structure learning for traffic prediction[J]. *Knowledge-Based Systems*, 2023, 278: 110885.
 - [30] Zhou K, Ziwei L, Yu Q, et al. Domain Generalization: A Survey[J]. *Ieee Transactions On Pattern Analysis and Machine Intelligence*, 2023, 45(4): 4396-4415.
 - [31] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP: International Conference on Machine Learning, 2019[C]. PMLR.
 - [32] Yuan Y, Chen Z, Wang Z, et al. Attention mechanism-based transfer learning model for day-ahead energy demand forecasting of shopping mall buildings[J]. *Energy*, 2023, 270: 126878.
 - [33] Park H, Park D Y, Noh B, et al. Stacking deep transfer learning for short-term cross building energy prediction with different seasonality and occupant schedule[J]. *Building and Environment*, 2022, 218: 109060.
 - [34] Himeur Y, Elnour M, Fadli F, et al. Next-generation energy systems for sustainable smart cities: Roles of transfer learning[J]. *Sustainable Cities and Society*, 2022, 85: 104059.
 - [35] Ribeiro M, Grolinger K, Elyamany H F, et al. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting[J]. *Energy and Buildings*, 2018, 165: 352-363.
 - [36] Fang X, Gong G, Li G, et al. A general multi-source ensemble transfer learning framework

- integrate of LSTM-DANN and similarity metric for building energy prediction[J]. *Energy and Buildings*, 2021, 252: 111435.
- [37] Lu H, Wu J, Ruan Y, et al. A multi-source transfer learning model based on LSTM and domain adaptation for building energy prediction[J]. *International Journal of Electrical Power & Energy Systems*, 2023, 149: 109024.
- [38] J. W, C. L, C. L, et al. Generalizing to Unseen Domains: A Survey on Domain Generalization[J]. *Ieee Transactions On Knowledge and Data Engineering*, 2023, 35(8): 8052-8072.
- [39] Qin X, Wang J, Chen Y, et al. Domain generalization for activity recognition via adaptive feature fusion[J]. *Acm Transactions On Intelligent Systems and Technology*, 2022, 14(1): 1-21.
- [40] Hu J, Qi L, Zhang J, et al. Domain generalization via Inter-domain Alignment and Intra-domain Expansion[J]. *Pattern Recognition*, 2024, 146: 110029.
- [41] Zhao C, Shen W. A domain generalization network combing invariance and specificity towards real-time intelligent fault diagnosis[J]. *Mechanical Systems and Signal Processing*, 2022, 173: 108990.
- [42] Gui S, Zhu Y, Qin X, et al. Learning multi-level domain invariant features for sketch re-identification[J]. *Neurocomputing*, 2020, 403: 294-303.
- [43] Du Y, Wang J, Feng W, et al. Adarnn: Adaptive learning and forecasting of time series: Proceedings of the 30th ACM international conference on information & knowledge management, 2021[C].
- [44] Wan A, Chang Q, Al-Bukhaiti K, et al. Short-term power load forecasting for combined heat and power using CNN-LSTM enhanced by attention mechanism[J]. *Energy*, 2023, 282: 128274.
- [45] Niu D, Yu M, Sun L, et al. Short-term multi-energy load forecasting for integrated energy systems based on CNN-BiGRU optimized by attention mechanism[J]. *Applied Energy*, 2022, 313: 118801.
- [46] Agga A, Abbou A, Labbadi M, et al. CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production[J]. *Electric Power Systems Research*, 2022, 208: 107908.
- [47] He D, Zhang H, Hao W, et al. An attention-based hybrid neural network for document modeling[J]. *Ieice Transactions On Information and Systems*, 2017, 100(6): 1372-1375.
- [48] J. Z, X. M, K. Z. Enhancing Attention-Based LSTM With Position Context for Aspect-Level Sentiment Classification[J]. *Ieee Access*, 2019, 7: 20462-20471.
- [49] Yan L, Wu Y, Yan L, et al. Encoder-Decoder Model for Forecast of PM2.5 Concentration per Hour: 2018 1st International Cognitive Cities Conference (IC3), 2018[C].2018.Aug.7-9.
- [50] Laubscher R. Time-series forecasting of coal-fired power plant reheater metal temperatures using encoder-decoder recurrent neural networks[J]. *Energy*, 2019, 189: 116187.
- [51] Guo J, Lin P, Zhang L, et al. Dynamic adaptive encoder-decoder deep learning networks for multivariate time series forecasting of building energy consumption[J]. *Applied Energy*, 2023, 350: 121803.
- [52] Li Q, Zhang X, Ma T, et al. A Multi-step ahead photovoltaic power forecasting model based on TimeGAN, Soft DTW-based K-medoids clustering, and a CNN-GRU hybrid neural network[J]. *Energy Reports*, 2022, 8: 10346-10362.

- [53] Iwana B K, Frinken V, Uchida S. DTW-NN: A novel neural network for time series recognition using dynamic alignment between inputs and weights[J]. Knowledge-Based Systems, 2020, 188: 104971.
- [54] Ye R, Dai Q. Implementing transfer learning across different datasets for time series forecasting[J]. Pattern Recognition, 2021, 109: 107617.
- [55] Yan N, Yao Y, Jia Z, et al. Online battery health diagnosis for electric vehicles based on DTW-XGBoost[J]. Energy Reports, 2022, 8: 121-128.
- [56] Sameen M I, Pradhan B, Lee S. Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment[J]. Catena, 2020, 186: 104249.
- [57] Sameen M I, Pradhan B, Lee S. Self-Learning Random Forests Model for Mapping Groundwater Yield in Data-Scarce Areas[J]. Natural Resources Research, 2019, 28(3): 757-775.
- [58] Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for hyper-parameter optimization[J]. Advances in Neural Information Processing Systems, 2011, 24.
- [59] Hoxha J, Codur M Y, Mustafaraj E, et al. Prediction of transportation energy demand in Turkiye using stacking ensemble models: Methodology and comparative analysis[J]. Applied Energy, 2023, 350: 121765.
- [60] Gao L, Liu T, Cao T, et al. Comparing deep learning models for multi energy vectors prediction on multiple types of building[J]. Applied Energy, 2021, 301: 117486.
- [61] Sankarananth S, Karthiga M, E. S, et al. AI-enabled metaheuristic optimization for predictive management of renewable energy production in smart grids[J]. Energy Reports, 2023, 10: 1299-1312.

Highlights:

- 1) Combining encoder and decoder architectures to build deep neural networks
- 2) Combining TDC and deep neural networks to train generalization model
- 3) Model training and testing using real office building energy consumption datasets
- 4) Setting up different comparison experiments on the source and target domains

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

--