Energy 270 (2023) 126878

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Attention mechanism-based transfer learning model for day-ahead energy demand forecasting of shopping mall buildings

Yue Yuan [a,b], Zhihua Chen [a,b], Zhe Wang [c,d], Yifu Sun [e], Yixing Chen [a,b,*]

[a] College of Civil Engineering, Hunan University, Changsha, 410082, China
[b] Key Laboratory of Building Safety and Energy Efficiency of Ministry of Education, Hunan University, Changsha, 410082, China
[c] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China
[d] Department of Civil and Environmental Engineering Employment, Hong Kong University of Science and Technology, Hong Kong
[e] Persagy Technology CO., Ltd., Beijing, 100096, China

ARTICLE INFO

Handling Editor: Henrik Lund

Keywords:
Transfer learning
Attention mechanisms
Shopping mall building
Energy prediction

ABSTRACT

The forecasting performance of data-driven models decreases rapidly with a limited training dataset. Herein, we sought to solve this problem by developing an attention mechanism-based transfer learning model and comparing its predictive ability in day-ahead energy consumption with those of three direct learning models: artificial neural networks with auto-regression (AR-ANN), random forest with auto-regression (AR-RF), and long short-term memory neural network (LSTM). Our target building was a large-scale shopping mall in Harbin, with 2 years of monitored data. The 2-months to 1-year data selected from the first year and all data from the second year were used as the training and testing sets, respectively. These models predicted the target building's peak electricity demand (PED) and total energy consumption (TEC). The results showed that the proposed transfer learning model outperformed the three direct learning models when data were insufficient in the training set. Specifically, the direct prediction models' lowest PED and TEC prediction errors were 34.34% and 26.32%, respectively, with 2-month training data available. In comparison, the corresponding prediction errors of the proposed model were only 12.48% and 10.78%, respectively. This study demonstrated the excellent performance of the proposed model with limited data.

1. Introduction

An accurate day-ahead energy prediction can play a vital role in building energy management, energy-saving potential evaluation [1], and greenhouse gas emission tracking [2]. Besides, building energy consumption prediction is also fundamental for the predictive control of building air-conditioning systems [3,4], individual control of electrical equipment [5,6] and integration of renewable energy systems [7]. Building energy consumption data are characterized by high dimensions, chaotic information, and unclear correlations [8]. For this reason, data-driven models have become a powerful technology in building energy consumption research [9] because of their simple modeling and resource-saving advantages.

The data-driven energy forecasting model can be considered a time-series forecasting problem [10]. Accordingly, the contribution of early studies in this field concentrated on developing and advancing time-series algorithms mainly. Because of the purse of higher accuracy, the algorithms developed from multiple linear regression (MLR) [11] and autoregressive integrated moving average (ARIMA) [12] to machine learning, even deep learning. Machine learning models have become more popular than statistical models for building prediction tasks, such as decision trees (DT) [13,14], support vector regression (SVR) [15], artificial neural networks (ANN) [16], random forest (RF) [17], convolution neural networks (CNN) [18], long short-term memory neural networks (LSTM) [19], and ensemble models [20–23]. Tian et al. [24] combined the EnergyPlus model and generative adversarial

Abbreviations: ACF, Autocorrelation coefficient; APE, Absolute percentage error; AR, Auto-regression; ARIMA, Autoregressive integrated moving average; CNN, Convolution neural networks; DT, Decision trees; GAN, Generative adversarial network; IQR, Interquartile range; LSTM, Long short-term memory neural network; MMD, Maximum mean discrepancy; MSE, Mean square error; NN, Artificial neural networks; NSFC, Natural Science Foundation of China; PED, Peak electricity demand; ReLU, Rectified linear unit; RF, Random forest; TEC, Total energy consumption.

* Corresponding author. College of Civil Engineering, Hunan University, Changsha, 410082, China.
E-mail address: yixingchen@hnu.edu.cn (Y. Chen).

https://doi.org/10.1016/j.energy.2023.126878
Received 5 July 2022; Received in revised form 16 December 2022; Accepted 4 February 2023
Available online 7 February 2023

network (GAN) to improve the prediction accuracy for buildings on a large scale. Fan et al. [10] proposed a deep generative modeling-based data augmentation strategy to improve short-term building energy predictions. Feng et al.[25] introduced introduced the uncertainty of using window shades in several machine learning energy prediction models, which can improve prediction efficiency without any complex simulation process. Nevertheless, although most models have an impressive performance in energy consumption prediction tasks of different buildings, few of them could be applied to other buildings directly because data-driven models are specifically designed, resulting in the need for sufficient historical data, which further increases the inconvenience of direct application.

Transfer learning [26] solves data limitations and reusable models. According to the definition, the kernel of transfer learning is used to reuse the knowledge acquired under sufficient data for an insufficient data task. Specifically, we typically define a domain with sufficient data as the source domain D_S and tasks performed on the source domain as the source domain tasks T_S. The domain and task we want to deal with but have insufficient data on will be defined as the target domain D_T and target task T_T [27]. The condition where $D_S \neq D_T$ and $T_S = T_T$ is called transductive transfer learning. The transfer learning scenario is the most active building energy forecasting research field. Transfer learning methods are generally divided into feature extraction and fine-tuning. Transfer learning based on feature extraction aims to reduce the difference between D_S and T_S by selecting good features through pre-training to improve T_T prediction performance on D_s, which directly learns by migrating parameters or prior distributions from D_S to D_T.

Nowadays, transfer learning has been widely applied to improve building performance in different research fields. As the most mature building performance research field involving artificial intelligence technology [28], there has been a large amount of state-of-the-art research on transfer-learning-based building energy consumption prediction. Houidi et al. [29] selected relevant and understandable features to build a transfer-learning model that can efficiently discriminate distinct home electrical appliances from energy-using profiles for residential buildings. Liu et al. [30] validated that transfer-learning-based strategies can be applied to detect, diagnose, and overcome data limitations for HVAC systems. Gao et al. [31] developed a transfer-learning model for indoor thermal comfort prediction tasks in multiple cities for a government office building. Owing to the limitations of experimental conditions, major studies are still building transfer learning models in educational buildings based on the Genome Project [32]. However, the potential for transfer learning across large-scale shopping malls has yet to be fully developed.

One big challenge in building energy forecasting is to make D_S and D_T as similar as possible in the transfer learning model. Therefore, many researchers have focused on selecting a suitable dataset for $D_{S.}$ Research in this area has shown that a model-based transfer-learning method can predict the thermal load for different residential buildings under the same energy station [33]. Ribeiro et al. [34] proposed a transfer learning method for cross-building energy consumption prediction based on seasonal and trend adjustments, which considers multiple time series features for multiple buildings. The results showed that using similar-domain datasets can improve educational buildings' energy consumption prediction accuracy. An ensemble tree-based transfer learning investigation was implemented based on a dataset from two leisure centers and an office building in Melbourne [35]. Tian et al. [36] proposed a similarity-based chained transfer learning model to take advantage of a well-trained model for educational buildings with insufficient data. Grubinger et al. [37] used a similar source domain to pre-train an energy forecasting model for residential buildings and validated climate control. These proposed methods of searching D_S based on D_T can promote transfer accuracy to a certain extent. However, there are diverse reasons for the differences between the source and target domains, such as building location, building climate region, and building function. Datasets of similar buildings can improve the accuracy of transfer learning. When there is no appropriate D_S, the transfer may not be effective and may sometimes result in a negative effect.

The other challenge is improving the antagonism and generalization ability of the prediction model. Fan et al. [38] compared several parameter-based structures to enhance building forecasting prediction and then analyzed how data availability and duration period availability influence parameter-based transfer learning. Fang et al. [39] proposed a building forecasting model with few labeled data to study the effects of different time horizons, architectures, and buildings, employing an LSTM as feature extraction and then fine-tuning a regression layer for domain adaption. Zhou et al. [40] proposed a novel approach to perform load prediction with no data or augmenting data in the case of a small dataset. Chen et al. [41] developed a hierarchical deep convolutional neural network based on transfer learning for fault identification in transformer rectifier units. These model-based approaches have high technical requirements and computational costs but have poor interpretability, which will limit the proliferation of model-based transfer-learning approaches. However, it remains to be seen whether the model can maintain high accuracy for different predictive tasks of a building.

In this case, we identified a major research gap: few researchers have focused on structuring transfer learning models based on the multi-source domain for energy prediction in commercial buildings. This study aims to develop an attention-based transfer learning strategy using the attention–CNN–LSTM method with the following steps: First, several pre-trained models were established with sufficient datasets in different source domains. Second, after the target domain was determined, the source domain was selected by calculating the similarity between the target domain and several source domains. Finally, we transferred the knowledge; the model was pre-trained from the source domain with a sufficient dataset to the target domain with insufficient data. This study sought to address the following questions.

1. At a minimum, how much data are needed to train a standalone building energy prediction model with satisfactory accuracy?
2. Can the attention mechanism be a powerful tool to predict the energy consumption for buildings with insufficient data by capturing the feature map from other buildings?
3. Is there an effective method to help select the domain source from several buildings?

2. Methodology

2.1. Outline

In this section, methods for constructing different models are introduced. The workflow of this study is shown in Fig. 1. The model construction process used in this study is as follows. Initially, the hourly energy consumption and meteorological data of large shopping mall buildings in four different climate zones were collected. These raw datasets were cleaned and preprocessed through outlier detection and missing-value filling. Next, the processed datasets were combined with time and holiday labels, outdoor weather conditions, and historical energy consumption values. The training and testing datasets were chronologically separated to maintain casualties and avoid information leakage. Finally, we developed an attention-based CNN-LSTM model, which was compared with three widely used load prediction algorithms: 1) autoregression deep neural network (AR-ANN), 2) autoregression random forest (AR-RF), and 3) long-short memory neural network (LSTM).

First, the four models were trained with 12-month data and then tested in the following year's dataset to determine whether these popular models could perform well. Next, we gradually reduced the available training dataset and evaluated the performance of each model using different amounts of data. The models were trained with 2-, 4-, 6-, 8-,

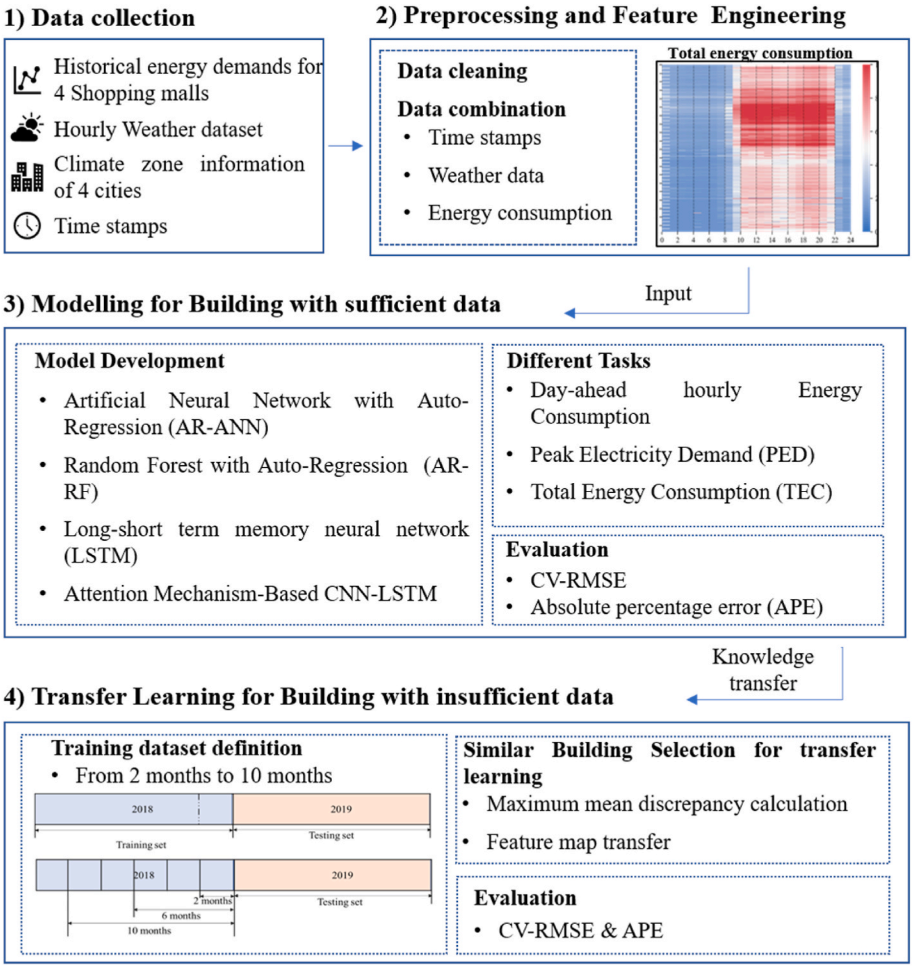

1) Data collection
Historical energy demands for 4 Shopping malls
Hourly Weather dataset
Climate zone information of 4 cities
Time stamps
2) Preprocessing and Feature Engineering
Data cleaning
Data combination
Time stamps
Weather data
Energy consumption
Total energy consumption
0 2 4 6 8 10 12 14 16 18 20 22 24
Input
3) Modelling for Building with sufficient data
Model Development
Artificial Neural Network with Auto-Regression (AR-ANN)
Random Forest with Auto-Regression (AR-RF)
Long-short term memory neural network (LSTM)
Attention Mechanism-Based CNN-LSTM
Different Tasks
Day-ahead hourly Energy Consumption
Peak Electricity Demand (PED)
Total Energy Consumption (TEC)
Evaluation
CV-RMSE
Absolute percentage error (APE)
Knowledge transfer
4) Transfer Learning for Building with insufficient data
Training dataset definition
From 2 months to 10 months
2018
2019
Training set
Testing set
2018
2019
2 months
6 months
10 months
Testing set
Similar Building Selection for transfer learning
Maximum mean discrepancy calculation
Feature map transfer
Evaluation
CV-RMSE & APE

Fig. 1. The workflow of this study.

and 10-month data, respectively, and then tested in the following whole-year dataset. Compared with the three standalone models, we proposed a transfer learning strategy in the attention-based CNN-LSTM model to select the source domain most similar to the target building from the three potential source domain buildings using the metrics of maximum mean discrepancy (MMD). Finally, we applied the same approach to different prediction tasks to determine the robustness of the proposed method.

2.2. *Data pre-processing and featurization*

2.2.1. *Data description and pre-processing*

The dataset was collected from four large-scale shopping malls with similar structures in different cities to validate the generalization ability of the proposed forecasting model in different climate regions. Our dataset was collected from buildings with the same function under the same brand, which means that the same operator operates these buildings. For this reason, buildings can be regarded as having similar energy-use patterns. Moreover, because these four buildings were designed and constructed by the same company, their energy use patterns were the same. When a new shopping mall is built, developers want to develop an energy-saving control strategy to achieve higher accuracy and lower data collection costs in the early operation stage. This study is dedicated to developing a novel method with transferability, high precision, high robustness, and high generalization capabilities for building energy prediction. For this reason, this study aims to determine when a new building starts operating and how its energy consumption can be predicted using a limited dataset.

According to the current building climate zoning standard (GB50178-1993) [42], China is divided into five climate zones (severe cold, cold, hot summer and cold winter, mild, hot summer, and warm winter) [43], three of which are covered in this study. The data quality statistics and basic information are presented in Table 1. According to Table 1, many missing values and extreme outliers in the original data are unfavorable for the prediction, so data preprocessing is required. In addition, both datasets provide hourly meteorological data, including the door temperature, humidity, and weather conditions (rain and sunny).

The dataset's quality directly affects the prediction model's performance [44]. Data cleaning and pre-processing are required to improve the reliability of the prediction model [45]. The pre-processing process in this study included 1) missing value processing, 2) extreme outlier processing, and 3) feature engineering.

(1) Missing value processing: Generally, missing values are handled in two ways: by deleting or filling them. Initially, processing variables with too many missing values should be deleted. After these variables were deleted, the linear interpolation method of adjacent values was used to fill in the missing values. In this study, linear interpolation was suitable when only a few points were missing because of the short time interval of the data collection stage.
(2) For extreme outliers, the boxplot method was used to mark the extreme outliers, and the linear interpolation method was used to replace them [46]. The definition of the boxplot in this study is Eq (1).

Table 1
Dataset quality overview.

Climate Zone	City	Duration	Energy consumption range (kWh)	Missing value number	Samples size
Cold	Beijing	2018.1.1–2019.12.31	[0,6172]	890	17,520
Hot summer and cold winter	Chengdu		[0,3580]	798	17,640
Hot summer and cold winter	Chongqing		[48,7005]	120	17,520
Severe Cold	Harbin		[37,7711]	120	17,520

$$L = Q_1 - 1.5 \times IQR \tag{1}$$

$$U = Q_3 + 1.5 \times IQR \tag{2}$$

where L represents the lower limit, and U represents the upper limit. IQR represents interquartile range. Extreme outliers were defined as values less than Q1-1.5IQR or greater than Q3+1.5IQR.

(3) Feature engineering: Selecting the most representative input variable from many features is necessary to reduce calculation costs and avoid model overfitting. This study used the autocorrelation coefficient (ACF) to determine the number of historical data input variables, such as the number of days that should be chosen to participate in the prediction model. In the equation, t represents the time series; N is the time series length, and k is the interval.

The peak and annual building load histograms are plotted in Fig. 2 and.3. Owing to the different climate zones, the energy consumption scales of each city are not the same. Harbin had the highest energy consumption level among the four cities, and Chongqing had the largest energy consumption range. In addition, large-scale duplicate variables must be manually deleted based on domain knowledge. The unit of the X-axis in Figs. 2 and 3 is the hourly energy consumption.

This original dataset records 34 types of energy consumption from various electrical equipment such as lights, elevators, and refrigeration. Here, we refer to them as end-use energy consumption. There are many missing values and extreme outliers in the original dataset, and the methods adopted in this study are time series and have strict continuity requirements. Therefore, the cleaning process of the original dataset used in this study was as follows.

(1) Repeated accumulated energy consumption variables were manually deleted through expert knowledge.
(2) End-use energy consumption with more than 100 missing items was deleted.
(3) Simple linear interpolation was performed for energy loss values to ensure data continuity.
(4) The box plot method eliminated energy consumption values beyond 1.5 times the quartile site.

2.2.2. Feature engineering

As shown in Fig. 4, the input data can be divided into three categories. The first two categories are historical energy consumption and historical meteorological parameters of buildings. The number of days to be considered was selected based on the ACF. The calculation function for the ACF is given by Eq. (3). In addition to determining the input of the historical energy consumption, because meteorological factors are added to the ANN, correlation analysis is first required to analyze outdoor variables. The inputs of the meteorological parameters and time labels were determined using Pearson correlation analysis. The Pearson correlation analysis function between the variables X and Y is shown in

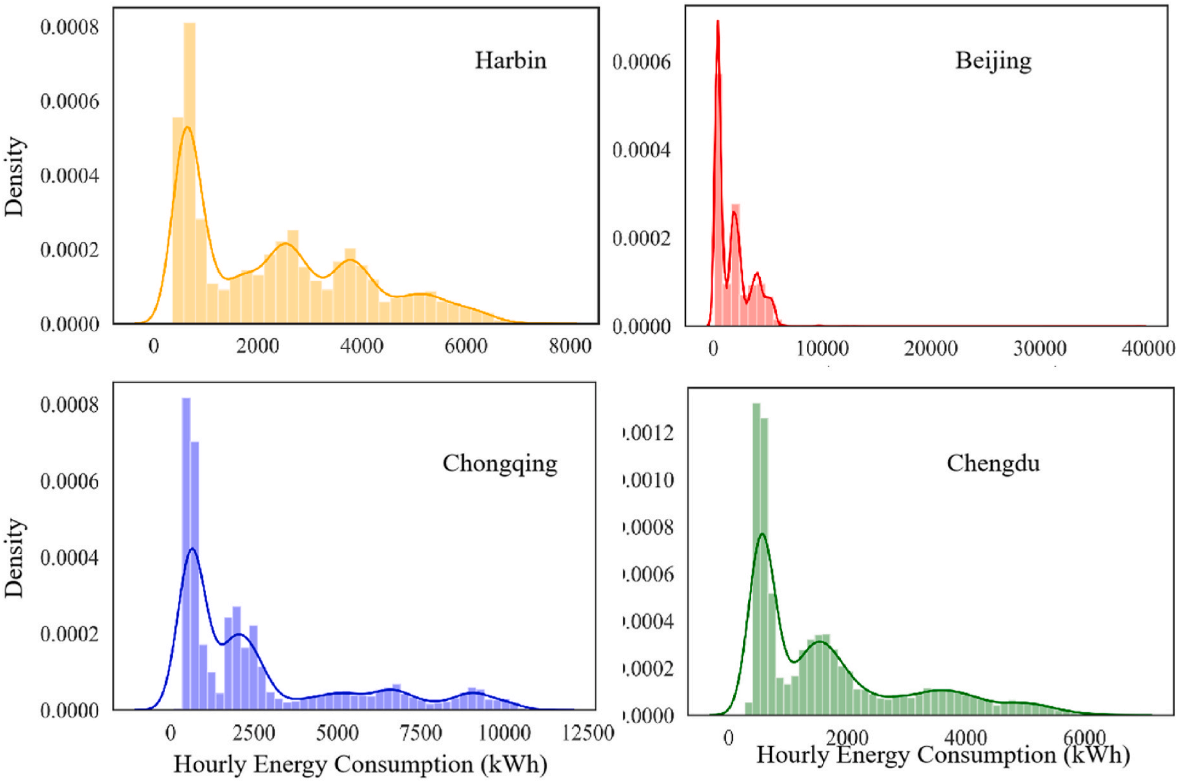


Fig. 2. Hourly electricity consumption distribution of the raw dataset of different cities.

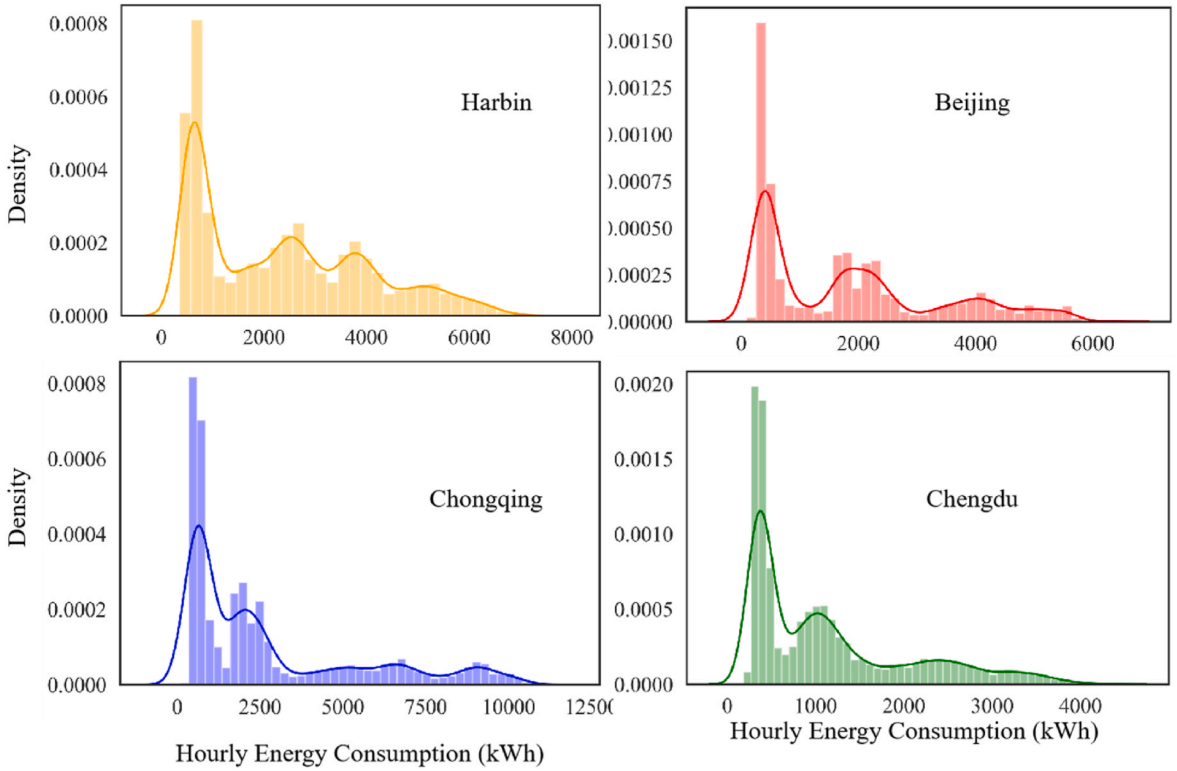


Fig. 3. Hourly electricity consumption distribution after data cleaning of different cities.

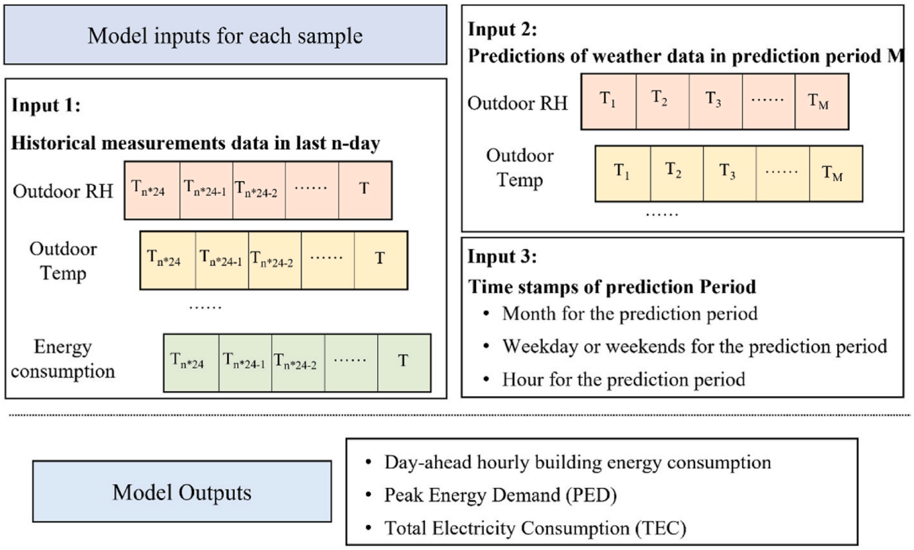

Model inputs for each sample
Input 1:
Historical measurements data in last n-day
Outdoor RH
T n*24
T n*24-1
T n*24-2
......
T
Outdoor Temp
T n*24
T n*24-1
T n*24-2
......
T
......
Energy consumption
T n*24
T n*24-1
T n*24-2
......
T
Input 2:
Predictions of weather data in prediction period M
Outdoor RH
T 1
T 2
T 3
......
T M
Outdoor Temp
T 1
T 2
T 3
......
T M
......
Input 3:
Time stamps of prediction Period
• Month for the prediction period
• Weekday or weekends for the prediction period
• Hour for the prediction period
Model Outputs
• Day-ahead hourly building energy consumption
• Peak Energy Demand (PED)
• Total Electricity Consumption (TEC)

Fig. 4. Overview of model inputs and outputs.

Eq. (4).

$$acf(k) = r_k = \frac{c_k}{c_0} = \frac{N}{N-k} \times \frac{\sum_{t=k+1}^{N}(x_t-\mu)(x_{t-k}-\mu)}{\sum\limits_{t=1}^{N}(x_t-\mu)(x_t-\mu)} \tag{3}$$

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2(Y-\overline{Y})^2}} \tag{4}$$

where, for the time series x_t; k is represented the interval, N is represented the length of the series, and μ is represented the average value of the time series.

Meteorological parameters were selected using Pearson correlation analysis. The second part is the outdoor meteorological conditions of the prediction period. The prediction task determines the length of the second part, and the variable selection is based on Pearson correlation analysis. The third part represents the time-related features, including the month (i.e., January to December), day type (i.e., workday or weekends), and time (0:00 to 23:00) of the forecast day.

In addition, the prediction task determination in this study was based on practical application scenarios. The accurate prediction of next-day energy use for buildings can benefit from renewable energy and energy storage technologies. For instance, hourly energy consumption guides the battery charging schedule. From the perspective of the ice storage system, peak electricity demand (PED) and total energy consumption (TEC) are sufficient. Accordingly, there are three prediction tasks based on different application scenarios: 24-h day-ahead energy consumption, daily PED, and daily TEC.

This study introduced autocorrelation and partial autocorrelation functions to determine each model's optimal number of days. According to the calculation results shown in Table 2, the previous 2, 3, 4, and 5 days are the inputs of Harbin, Beijing, Chongqing, and Chengdu, respectively.

Correlation analysis was performed to determine the input variables from the outdoor environment parameters. Meteorological parameters and time labels with the greatest correlation with energy consumption were obtained. The results are presented in Table 3. This variable was not included in this dataset. To reduce the complexity of the model and improve the calculation speed, only the variables whose correlation absolute value is greater than 0.2 are retained, that is, outdoor temperature, atmospheric pressure, and visibility or Air Quality Index （AQI） for some cities. Correlation analysis aims to select the most relevant features for all features. Thus, the computation and risk of overfitting were reduced.

This study also analyzed different time labels, and the results are shown in Table 4. The results demonstrate that the year, month, week, and energy consumption have a low correlation. Whether it is a holiday has a significant impact on energy consumption and should be added as an input variable to the prediction model.

2.3. Data-driven algorithm modeling

2.3.1. AR-ANN model

ANNs, deep-learning frameworks, are the most widely used artificial intelligence algorithms [47]. The ANN structure consists of many similar biological neural network processing units with interconnected nonlinear network structures, including the input, multiple hidden, and output layers. Nonlinear elements were introduced into the network through different activation functions to solve complex problems. In a study by Wang et al. [48], the ANN model was proven to accurately estimate nonlinear problems such as energy consumption forecasting.

Although ANNs exhibit powerful performance in nonlinear fitting problems, they cannot obtain temporal information from time series. As shown in Fig. 2, the input structure should be reshaped first instead of directly training the dataset to use historical data. Thus, historical information from the original dataset can be further used in the ANN model, and a higher forecasting precision can be obtained. This process is called auto-regression (AR). Accordingly, the AR-ANN model with a reshaped input dataset was named the AR-ANN model.

2.3.2. AR-RF model

RF is a popular artificial intelligence algorithm, in addition to ANNs [49]. Like a forest, its basic principle is an algorithm that integrates multiple trees through the bagging idea of ensemble learning: its basic unit is the decision tree. The predictions are made by averaging the predictions for each decision tree.

In this study, the CART regression tree was selected as the weak classifier, and the bagging algorithm was used for integration. The CART regression tree adopts the minimum mean square error (MSE) for error correction; that is, for datasets D_1 and D_2 divided into both sides of the corresponding arbitrary partition points for any partition feature A, the corresponding feature and eigenvalue partition points that minimize the mean square errors of D_1 and D_2, respectively, and the sum of the mean square errors of D_1 and D_2 are obtained [50]. The equation is as follows:

$$\underbrace{min}_{A,s}\left[\underbrace{min}_{c_1}\sum_{x_i,\in D_1(A,s)}(y_i-c_1)^2+\underbrace{min}_{c_2}\sum_{x_i,\in D_2(A,s)}(y_i-c_2)^2\right] \tag{5}$$

Similar to AR-ANN, the processing of the input data in RF is also static, with no time-series traits. Therefore, the input for RF in this study also needs to be reshaped beforehand. Therefore, the model applied in this study was also called AR-RF.

Table 2
Historic time input specification.

Cities	Previous day's input to models
Harbin	Last 2 days
Beijing	Last 3 days
Chongqing	Last 5 days
Chengdu	Last 4 days

2.3.3. LSTM model

LSTM is an advanced, recurrent neural network model with feedback connections. It can process not only single data points (such as images) but also entire sequences of data [51]. There are two reasons for choosing LSTM to develop the prediction model in this study:1) LSTM is sensitive to historical states and can process dynamic information; 2) it performs a multivariate nonlinear prediction task. The LSTM neuron structure is shown in Fig. 5, where there are three door structures in the neuron structure: the input gate, output gate, and forget gate. In LSTM, the first step is to determine the discarded information in the cell status through the forgotten door. The second step is to determine the information that must be placed in the cells in the input gate. The third step is to set the output value in the output door.

2.4. Maximum mean discrepancy

MMD is the most popular method for domain adaptation in transfer learning. It measures the distance between two different but related distributions.

Based on the samples with two distributions, the mean value of the function f for the samples with different distributions was obtained by looking for the continuous function f in the sample space. The mean discrepancy between the two distributions corresponding to F was obtained by determining the difference between the two means. An F that maximizes favor was selected, and MMD was obtained. The MMD is the test statistic used to determine whether the two distributions are identical. Its calculation formula is defined in Eq. (5):

$$MMD(F,p,q)=\sup_{\|f\|_{H\leq1}}E_p[f(x)]-E_q[f(y)] \tag{6}$$

where the distribution of x is p and y is q. Sup represents the upper bound. Eq. (6) represents the expectation of P, and f represents the mapping function. $H\leq 1$ means that function f in the regenerated Hilbert space should equal 1.

2.5. Attention–CNN–based pre-trained model

The proposed Attention–CNN–LSTM model can capture long-distance information and comprehensively mine information from time-series data. The CNN module was optimized by the attention mechanism adopted to extract the useful feature map from the pre-trained model. Subsequently, the LSTM can transfer the feature map to the target source and complete the forecasting task. Moreover, the proposed model can handle various forecasting tasks. Because this study aimed to build a day-ahead model that can be used for an entire year, the long-term memory capacity of LSTM to store feature information was essential. Furthermore, the convolutional block attention model highlighted the importance of the data features at different times to improve the model's performance.

2.5.1. Basic concept of the convolution neural network

Generally, the convolution neural network includes convolution,

Table 3
Correlation analysis between outdoor parameters and energy consumption.

Cities	Pressure	Wind direct degree	Outdoor temperature	Weather	Wind direction	Wind speed	Wind direct d	Visibility	Humidity
Beijing	−0.33	0.07	0.56	−0.02	−0.07	0.23	0.17	0.09	−0.12
Chengdu	**−0.44**	0.02	**0.64**	−0.01	−0.01	0.09	0.01	**0.36**	−0.31
Chongqing	**−0.44**	−0.13	**0.6**	0.01	−0.02	0.04	−0.01	**0.35**	**−0.41**
Harbin	−0.05	−0.01	**0.25**	0.01	0.02	0.14	0.01	0.14	−0.17

Table 4
Correlation analysis between time labels and energy consumption.

	Year	Month	Week	Weekday	Holiday
Beijing	−0.06	0.06	0.01	0.01	0.21
Chengdu	−0.04	0.11	0.04	0.03	0.15
Chongqing	0.03	−0.02	0.01	0.02	0.34
Harbin	0.06	0.1	0.01	0.1	0.23

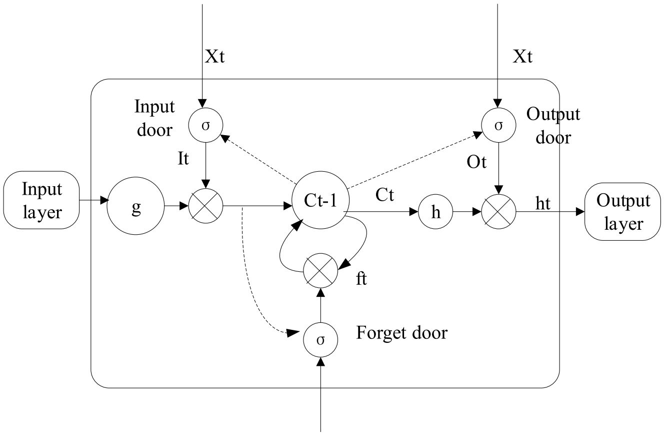


Fig. 5. LSTM schematic diagram.

pooling, and fully connected layers. Each convolution layer is composed of several convolution units, and the backpropagation algorithm optimizes the parameters of each convolution unit. The purpose of the convolutional layer is to extract different input features. The first convolution layer can only extract some low-level features, such as edges, lines, and angles. By contrast, a network with more layers can iteratively extract more complex features from low-level features.

The pooling layer reduces the shape of the input matrix and extracts the features. There are several different pooling methods in the pooling layer, among which average pooling is the most popular method that calculates the average value for each matrix pooling area. Another method is max pooling, which calculates the maximum value for each matrix-pooling area.

2.5.2. Attention mechanism-based CNN module

Inspired by human vision, experts and scholars have proposed an attention mechanism to efficiently allocate information processing resources, widely applied in image recognition, semantic segmentation (NLP) [52], and other fields.

The principle of the attention-CNN module based on the attention mechanism is shown in Fig. 6, which includes three parts: 1) the data-process layer, 2) the attention-CNN layer, and 3) the output layer. The attention mechanism is implemented by retaining the intermediate outputs of the LSTM encoder on the input sequence and then training a model to selectively learn these inputs and the associated LSTM encoder sequences to calculate the model outputs. In other words, the probability of generating each item in the output sequence depends on which items are selected in the input sequence. The training dataset was segmented into l and s ($l < s$). Each CNN cell input length l; the corresponding attention mechanism module inputs a unit of s. Because the s units are longer than the l units, the attention mechanism can obtain more comprehensive information than the CNN.

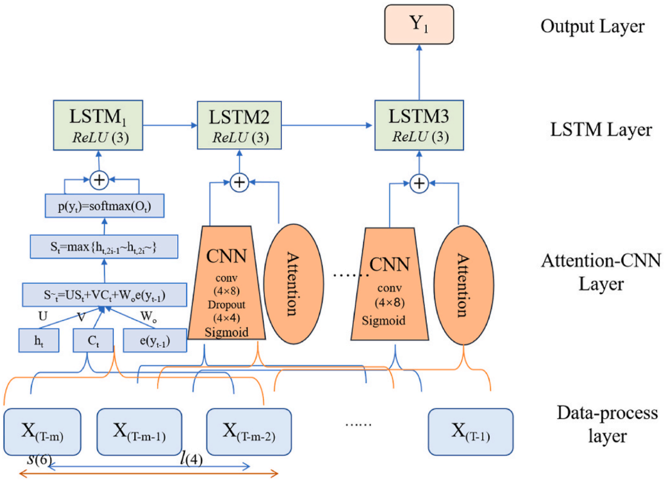


Fig. 6. Attention–CNN–LSTM schematic diagram.

Specifically, the attention mechanism-based CNN model is based on the convolution block attention module (CBAM) [53] and is introduced into the energy consumption prediction model to deal with the significant difference in short sequence features ignored by existing structures and extract significant fine-grained features. In addition, the model can obtain the time dependence more effectively. If convolution unit F is represented by Eq. (7), the process of the channel attention module and spatial attention module is represented by Eq. (8) and Eq. (9), respectively:

$$F \in R^{C\times H\times W} \tag{7}$$

$$F' = M_C(F) \otimes F \tag{8}$$

$$F'' = M_S(F') \otimes F' \tag{9}$$

where F represents the input feature; F' and F'' represent the channel-refined features; C denotes channel attention, and S denotes spatial attention.

CBAM consists of channel and spatial attention modules, the workflow shown in Fig. 7. After the first convolution layer, the input features become several convolutional units. First, the convolution unit is input to the channel attention module and conducts average and max pooling. The input unit becomes a one-dimensional vector element-wise through a fully connected layer. As shown in Eq. (10), this step aggregates the spatial information of the feature mapping and compresses the input feature map to generate channel attention. Second, a one-dimensional vector is an input to the spatial attention module. We conducted max pooling and average pooling again for the one-dimensional vector. Finally, the concatenation of these two feature maps and the production of the final map is shown in Eq. (11).

$$\begin{aligned} M_C(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^C_{avg}\right)\right) + W_1\left(W_0\left(F^C_{\max}\right)\right)\right) \end{aligned} \tag{10}$$

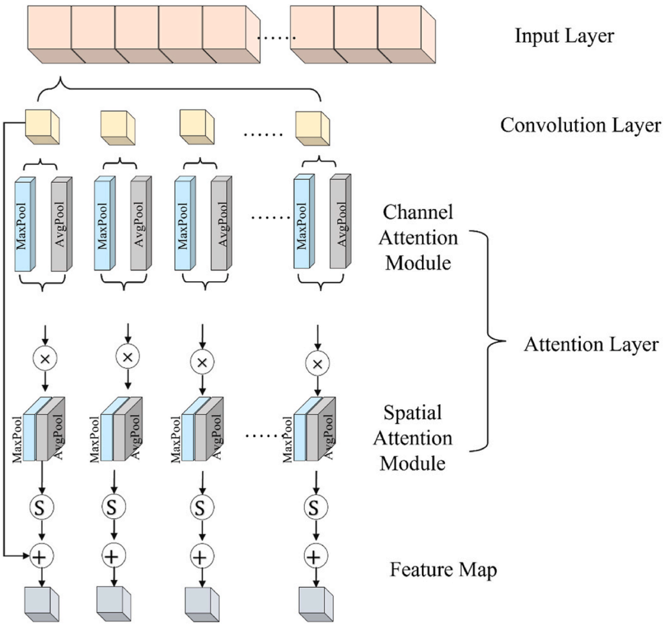


Fig. 7. Arrangement of attention modules.

$$M_s(F)=\sigma\left(f^{7\times7}\left(AvgPool(F);M=\sigma\left(f^{7\times7}\left(\left[F^S_{avg};F^S_{\max}\right]\right)\right)\right)\right) \tag{11}$$

where *MLP* represents a fully connected layer, and *AvgPool* and *MaxPool* represent the average-pooling and max-pooling conduction, respectively. *W* is the weight of *F* in the pooling layer.

In addition, to avoid overfitting, the input of each parallel module had adjustable overlapping parts. Adjusting the overlapping step size of the input units can make the model better adapt to data with different principles and further expand the length of the overall input of the joint model to capture more accurate long-term features. The inputs between the parallel modules do not overlap significantly.

2.6. Grid search of hyperparametric optimization

In a machine learning study, some parameters will be tuned with manual methods to achieve a better performance model. This study used a grid search [47] method to obtain the best hyperparameters for each model. The grid search included three steps: 1) optimization of the value range of the parameters, 2) calculation of the combination of all parameters, and 3) cross-validation for determining the best combination.

This study applied the grid optimization method to optimize the ANN, RF, and LSTM structures. Only three essential parameters were optimized for each model to reduce computation costs in model optimization. And the range of the parameters was selected according to Refs. [10,39]. In principle, the sum of the multiple of the number of nodes and the number of hidden layers should be smaller than the sample size. The specific grid search settings for the three model optimizations are summarized in Table 5.

2.7. Description of the four different models

To demonstrate the superiority of our proposed method (attention-mechanism CNN and LSTM, referred to as At–CNN–LSTM), their widely used building load prediction algorithms were selected as benchmark models: artificial neural network combined with auto-regression (AR-ANN), random forest combined with auto-regression (AR-RF), and LSTM (see Table 6) In addition, the processor for computation in this study was a 2.90 GHz Intel Core i7-10700 F. All computations were conducted in Python 3.8 using the neural network construction package TensorFlow.

Table 5
The grid-search settings for optimization of each model.

Models	Parameters	Grid-search values
ANN	The number of hidden layers	1, 2, 3
	The number of nodes for each hidden layer	16, 32, 64
	The activation functions in hidden layers	ReLU, Sigmoid, Tanh
Random Forest	The maximum number of iterations of the weak learner	100, 200,300
	The maximum depth of the weak learner	10, 20, 50
	The maximum number of features considered in the weak learning	3, 5, 6
LSTM	The number of hidden layers	1, 2, 3
	The number of nodes for each hidden layer	16, 32, 64
	The number of memory cells	Finer, 10, 100

Table 6
Four building load prediction algorithms to be compared.

Model	Specific instructions
AR-ANN	AR is initially used to process historical data. Next, by combining meteorological and historical data, the ANN algorithm performs forecasting tasks.
AR-RF	AR is initially used to process historical data. Next, by combining meteorological and historical data, the RF algorithm performs forecasting tasks.
LSTM	LSTM algorithm is applied for the data only
At–CNN–LSTM	By combining meteorological and historical data as input datasets, the forecasting tasks are conducted using Attention–CNN–LSTM.

2.8. Evaluation

The evaluation criteria selected in this study were based on the ASHRAE Guideline 14 [54]. The variation coefficient of the root means a square error of variance (CV-RMSE) was used to evaluate the performance of the models. According to ASHRAE, the CV-RMSE of the hourly data simulation value should be less than 30%. The formula is shown in Eq. (12).

$$CV-RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2}{n}}\Bigg/\frac{\sum_{i=1}^{n}Y_i}{n} \tag{12}$$

where *n* represents the number of variable *Y*, and *i* represent the order.

In addition, the absolute percentage error (APE) was introduced to evaluate the deviation between the measured and predicted values, as shown in Eq. (13):

$$APE=\left|\frac{\widehat{y}_i-y_i}{y_i}\right|\times100\% \tag{13}$$

where *y represents the series, and* represent the order.

Furthermore, the Friedman test was used to compare the evaluated models. The formula is shown in Eq. (14) and Eq. (15).

$$\tau_{\chi^2}=\frac{12N}{k(k-1)}\left(\sum_{i=1}^{k}r_i^2-\frac{k(k+1)^2}{4}\right) \tag{14}$$

$$\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} \tag{15}$$

In this function, suppose we compare *k* algorithms on *N* datasets and let r_i denote the average order value of the *i*th algorithm. The mean and variance of r_i are (k+1)/2 and (k^2-1)/12*N*, respectively.

2.9. Test cases

As shown in Fig. 8, we first selected the whole-year data as the

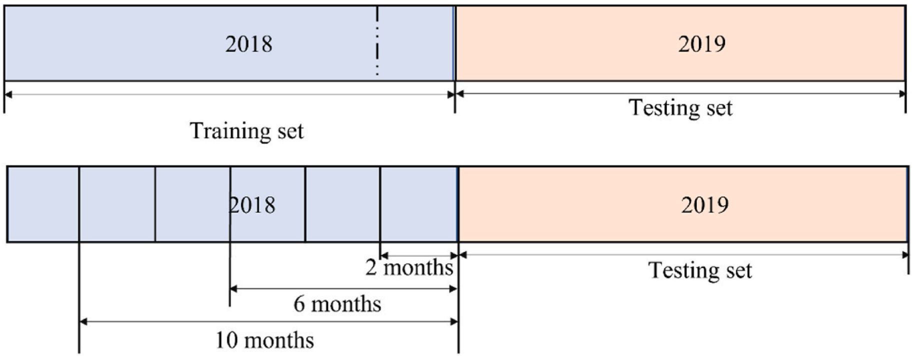


Fig. 8. The training set and test set.

training set under data-sufficient conditions, and the validation set was randomly selected from the training set. The following-year data were selected as the test set. Subsequently, we successfully reduced the amount of data available for training. In this study, 2-, 4-, 6-, 8-, and 10-month data were obtained from the target building. The prediction predicted the day-ahead PED and day-ahead TEC for 2019. For instance, the 2-month training dataset included November 1, 2018, to December 31, 2018. It is worth mentioning that, in the process of practical data collection, we could not obtain the ideal collection duration. To facilitate the presentation of the predicted results, we fixed the testing dataset and changed the collection time of the training set from 2 months to 10 months.

2.10. *Summarize the proposed method*

Fig. 9 summarizes the process given above: the pre-trained model of source domain building is developed with a sufficient dataset first. The source building is considered to be similar because of the function and management strategy, and the target task is the same. And the MMD is adopted to select the optimal source domain from several buildings. Finally, the prediction model of the target task is developed with the following steps: the model structure and feature map, the same as the pre-trained model, are first set up; the parameters of the pre-trained model are used as the initialization parameters of the target task model. The performance of this framework is evaluated from two aspects involving prediction error and stability under different training sample sizes of target tasks.

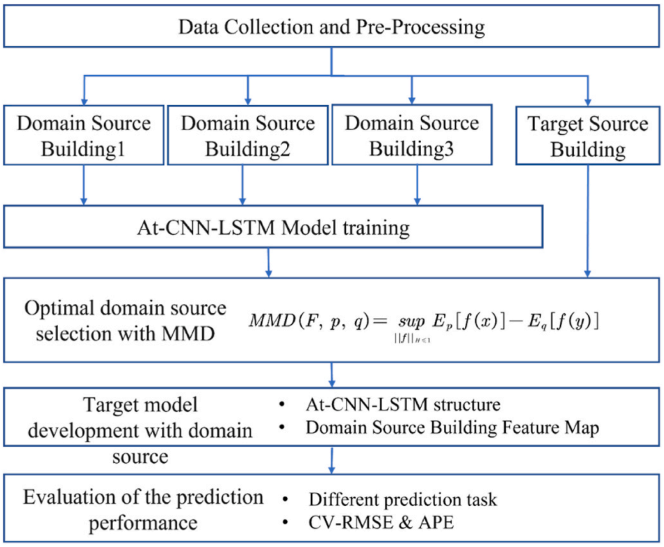

Data Collection and Pre-Processing
Domain Source Building1
Domain Source Building2
Domain Source Building3
Target Source Building
At-CNN-LSTM Model training
Optimal domain source selection with MMD
$MMD(F, p, q) = \sup_{\|f\|_{H<1}} E_p[f(x)] - E_q[f(y)]$
Target model development with domain source
• At-CNN-LSTM structure
• Domain Source Building Feature Map
Evaluation of the prediction performance
• Different prediction task
• CV-RMSE & APE

Fig. 9. The block diagram of the proposed model.

3. Result analysis

3.1. *Model performance without transfer learning*

In this section, the performance of the proposed model was comprehensively compared. We randomly divided the first-year (2018) data into prediction and training sets and used the second-year (2019) data as a validation model. These models were trained well, and no under-fitting or over-fitting phenomenon was observed; the accuracies of the training, testing and validation sets were elucidated. It can be seen that the prediction performance has little difference between each model with one-year training data set. It indicates that the algorithm will not be the major challenge for prediction task under sufficient data condition.

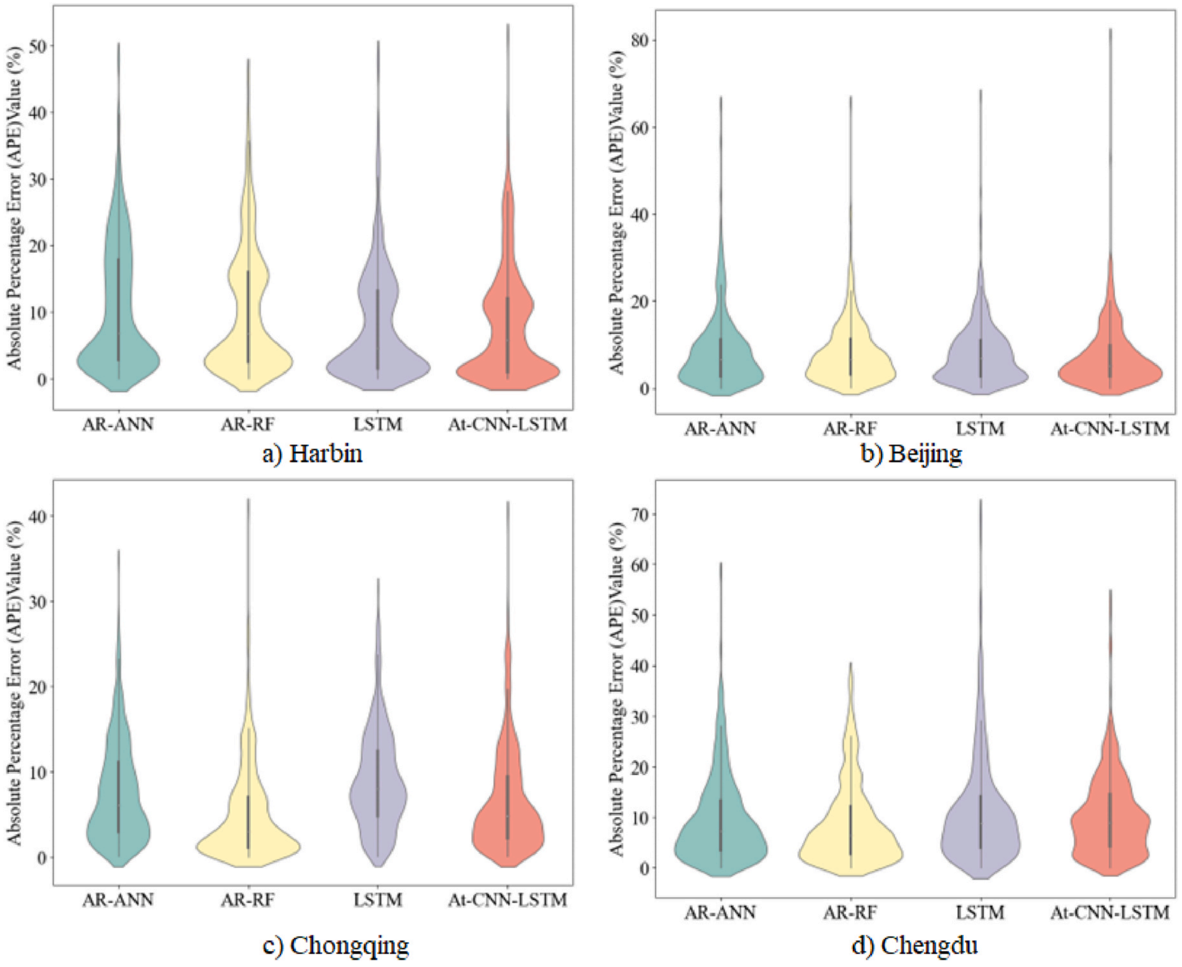
3.1.1. *Result evaluation for multi-output prediction tasks*

This study defined multi-output prediction tasks, namely 24-h day-ahead prediction. For the 24-h day-ahead prediction task, the best CV-RMSE was achieved at 6.49%, compared with AR-ANN, AR-RF, and LSTM, showing improvements of 2.71%, 0.98%, and 2.89%, respectively. The maximum CV-RMSE was reduced by approximately 6.90% in the 24-h day-ahead prediction tasks, and the average improvement is 2.14% which is more significant for a model with a poor prediction effect. As described in Table .7, the CV-RMSE results of the training, testing and validation datasets for all the models met the ASHRAE guideline. The AR-ANN, AR-RF, and LSTM showed prediction error values of 9.20%–15.2%, 7.47%–15.14%, and 9.38%–19.39%, respectively, and the prediction error range of At–CNN–LSTM was 6.49%–12.49%. This illustrates that the model developed in this study can deliver accurate building load prediction in different climate zones.

Furthermore, Fig. 10 illustrates the stability of each model in the 24-

Table 7
Model comparison in terms of prediction accuracy (CV-RMSE) for multi-output tasks.

		Training set	Testing set	Validation set
Harbin	AR-ANN	2.98%	7.38%	15.20%
	AR-RF	2.71%	6.21%	15.14%
	LSTM	2.71%	6.01%	19.39%
	At–CNN–LSTM	2.71%	6.01%	12.49%
Beijing	AR-ANN	3.00%	8.21%	9.20%
	AR-RF	2.39%	6.48%	7.47%
	LSTM	1.57%	6.48%	9.38%
	At–CNN–LSTM	1.23%	2.65%	**6.49%**
Chongqing	AR-ANN	6.51%	7.00%	11.00%
	AR-RF	2.63%	6.10%	10.88%
	LSTM	1.72%	6.53%	10.65%
	At–CNN–LSTM	0.17%	6.31%	**9.69%**
Chengdu	AR-ANN	5.54%	10.36%	10.69%
	AR-RF	3.38%	8.82%	8.32%
	LSTM	2.98%	5.36%	9.44%
	At–CNN–LSTM	2.98%	5.36%	8.41%



a) Harbin

b) Beijing

c) Chongqing

d) Chengdu

Fig. 10. Absolute percentage error (APE) distribution in different cities.

h day-ahead prediction task. The results showed that the APE distributions of all models conformed to the Gaussian distribution, indicating the effectiveness of the models. In addition, compared with AR-ANN, AR-RF, and LSTM, the attention mechanism can improve the prediction stability for the algorithms.

Table 8
Parameter configuration of the compared model structure.

Algorithm	Task	hyperparameter	Harbin	Beijing	Chengdu	Chongqing
ANN	PEC	layer	2	1	2	2
		Unit of each layer	16	16	16	16
		activation functions	*ReLU*	*ReLU*	*ReLU*	*ReLU*
	TEC	layer	2	1	2	2
		Unit of each layer	16	16	16	16
		activation functions	ReLU	ReLU	ReLU	ReLU
RF	PEC	weak learner	100	200	100	200
		maximum depth	20	30	20	20
		maximum number of features	5	3	5	3
	TEC	weak learner	100	100	200	200
		maximum depth	20	50	30	20
		maximum number of features	5	3	5	3
LSTM	PEC	layer	2	2	1	2
		Unit of each layer	16	8	16	16
		activation functions	10	Finer	Finer	Finer
	TEC	layer	2	2	1	2
		Unit of each layer	8	8	16	16
		number of memory cells	10	Finer	Finer	Finer
At–CNN–LSTM	PEC	layer	1	1	1	1
		Unit of each layer	5	5	6	5
		activation functions	Finer	Finer	Finer	Finer
	TEC	layer	1	1	1	1
		Unit of each layer	5	8	5	5
		number of memory cells	Finer	Finer	10	Finer

3.1.2. PED and TEC prediction result

Table 8 shows the detailed hyperparameter setting for all models after the grid search. In order to prevent overfitting, based on the empirical value, the parameters should not exceed one third of the number of training sets.

The prediction results of the TEC and PED models for different cities are presented in Table 9 to demonstrate the effectiveness of the proposed model. Similar to the data in Table 9, we selected 2018 as the training set and 2019 as the testing data. In addition, 20% of the data were randomly used as a validation set in the training set. For PED and TEC predictions, the smallest CV-RMSE values of the proposed model were 8.44% and 7.21%, respectively, and the largest CV-RMSE differences were 8.73% and 9.61%, respectively.

Table 10 shows the Friedman test results for each algorithm for the different city datasets. A P value less than 0.05 indicates a significant difference between algorithms; a P value greater than 0.05 indicates no significant difference between algorithms. The results showed that the algorithms differed when the amount of data was small.

According to the data sufficiency conditions, Harbin *had* the worst performance in the TEC and PED forecasting tasks. This was caused by the large amount of data with noise and error, which are difficult to be cleared only by pre-processing. Therefore, we selected Harbin as the target building and the other three cities as domain sources to demonstrate the efficiency of the transfer-learning approach. First, the sample sizes of the training sets were gradually increased from 2 to 10 months, with the MMD between the domain source and target source. Subsequently, the prediction performances of the three algorithms and three different domain source transfer learning models were discussed. Under data-poor conditions, 72 models were established.

3.1.3. Maximum mean discrepancy results for different cities

This study used the MMD method to calculate the distribution similarity of existing datasets in the source and target domains to select the source domain that is most similar to the target domain. The results are presented in Table 11. According to the results, the similarity between Ds and Ts decreased as the amount of data increased. In addition, the calculation results showed that the correlation between the energy consumption distribution and climate region was insignificant. The maximum mean discrepancy evaluates the dataset most similar to the target domain (Harbin).

3.1.4. Experimental results for transfer learning model

After the grid-search, the attention–CNN–based model used a 1D convolutional layer to extract the local temporal features from the input layer. The number of filters was 64, with a kernel size of 4 and a rectified linear unit (*ReLU*) activation function. A dropout layer followed the convolutional operations and was connected to a bidirectional recurrent layer with 128 LSTM units. The activation function used was *tanh*. Subsequently, another dropout layer was adopted. The attention module was connected to the dropout layer. Finally, the multi-output dense layer was the final output structure. And the detail were shown in Table 12.

Table 10
The Friedman test result for PED and TEC.

	TEC	PED
Friedman chi-squared	9.3	2.7
P value	0.255	0.44

Performance evaluation for transfer learning tasks.

Table 11
The Maximum mean discrepancy result of Harbin.

	2 months	4 months	6 months	8 months	10 months
Beijing	**0.68**	**0.57**	**0.52**	**0.46**	**0.44**
Chongqing	0.65	0.56	0.5	0.43	0.43
Chengdu	0.57	0.50	0.45	0.41	0.39

Table 13 shows that the prediction accuracy increased significantly with data availability. Table 13 summarizes the prediction accuracy for different cities. A decreasing trend in the CV-RMSE value was observed with increased data availability. Until the scope of the training set can cover the cooling season, the model's accuracy will reach a useable range. However, when the amount of data is sufficient, transfer learning is sometimes invalidated, producing a negative transfer effect. In

Table 12
Parameters of attention–CNN–LSTM

Input (365 × 4× 6)	
Attention (4 × 6)	CNN (4 × 5)
Conv (4 × 5)	
Drop Out (4 × 5)	Conv (4 × 5)
Avepool (2)	Conv (4 × 5)
Conv (4 × 8)	Conv (4 × 5)
Conv (4 × 8)	Maxpool (2)
Maxpool (2)	
attention_vec (Permute) (4 × 8)	
Sigmoid	
multiply_5 (Multiply) (4 × 8)	
Maxpool (2)	
MLP-32	
lSTM(3)	
LSTM (1)	

Table 9
Model comparison in terms of prediction accuracy (CV-RMSE) for PED and TEC.

		PED			TEC		
		Training set	Testing set	Validation set	Training set	Testing set	Validation set
Harbin	AR-ANN	5.38%	11.70%	11.43%	4.63%	7.35%	9.19%
	AR-RF	3.28%	9.34%	9.25%	2.23%	5.97%	9.88%
	LSTM	5.49%	7.95%	12.16%	4.87%	7.91%	8.75%
	At–CNN–LSTM	4.55%	8.26%	**9.48%**	3.46%	6.28%	**7.21%**
Beijing	AR-ANN	5.22%	10.89%	11.62%	5.40%	7.30%	12.31%
	AR-RF	3.28%	9.34%	12.25%	2.61%	6.65%	13.54%
	LSTM	3.46%	8.23%	12.16%	4.25%	5.84%	7.21%
	At–CNN–LSTM	5.20%	9.47%	**8.44%**	4.97%	6.25%	**10.29%**
Chengdu	AR-ANN	9.90%	10.17%	13.81%	9.25%	13.96%	12.48%
	AR-RF	6.03%	13.41%	16.90%	5.00%	10.52%	11.50%
	LSTM	4.35%	8.84%	15.36%	7.46%	13.39%	9.98%
	At–CNN–LSTM	4.55%	8.26%	**8.42%**	5.05%	6.28%	**8.71%**
Chongqing	AR-ANN	7.52%	13.78%	15.72%	6.48%	8.95%	17.72%
	AR-RF	4.79%	10.06%	17.47%	4.03%	7.52%	13.69%
	LSTM	6.55%	9.49%	16.88%	4.64%	6.47%	9.26%
	At–CNN–LSTM	6.33%	6.22%	**8.74%**	5.56%	5.54%	**8.11%**

Table 13
Summarize PED and TEC prediction error (CV-RMSE).

Task	Algorithm/Domain source	Training dataset scale				
		2 months	4 months	6 months	8 months	10 months
PED	AR-ANN	33.77%	21.16%	16.40%	14.59%	13.78%
	AR-RF	34.34%	32.03%	13.18%	12.58%	15.72%
	LSTM	32.35%	34.72%	18.68%	16.23%	15.62%
	At–CNN–LSTM-Beijing	**12.48%**	**14.33%**	**13.55%**	**12.87%**	**12.45%**
	At–CNN–LSTM-Chongqing	12.84%	15.47%	14.50%	13.27%	13.16%
	At–CNN–LSTM-Chengdu	13.95%	15.56%	14.18%	12.78%	12.97%
TEC	AR-ANN	30.45%	24.17%	20.52%	16.45%	12.51%
	AR-RF	26.32%	24.32%	20.90%	12.32%	10.23%
	LSTM	23.28%	21.29%	14.88%	14.36%	9.67%
	At–CNN–LSTM-Beijing	**10.78%**	**11.32%**	**12.22%**	**12.53%**	**8.35%**
	At–CNN–LSTM-Chongqing	13.74%	14.23%	12.63%	15.23%	13.23%
	At–CNN–LSTM-Chengdu	13.85%	13.94%	12.35%	12.32%	14.28%

conclusion, when the data availability is insufficient to cover the cooling season, transferring feature maps distributed from similar energy profiles can effectively improve the usability of the model.

In addition, the model trained from the source domain with the highest value can be transferred to the target domain with the highest accuracy, indicating that the MMD is an indicative metric for selecting the most transferrable source domain for transfer-learning tasks. Our proposed model's minimum prediction error (CV-RMSE) is 12.48% in PED prediction tasks and 10.78% in TEC prediction tasks with 2-month data available. The corresponding minimum prediction error (CV-RMSE) is 14.33% in PED prediction tasks and 11.32% in TEC prediction tasks with 4-month data available. The differences between the transfer learning and standalone machine learning models ranged from 18.40% to 21.86% in PED prediction tasks and 9.43%–19.67% in TEC prediction tasks using 2-month data. The difference ranged from 5.60% to 20.39% in PED prediction tasks and from 7.06% to 12.85% in TEC prediction tasks using 4-month data. In addition, compared with that with 12-month data, the accuracy of training sets with 8- to 10-month data exhibited a slight fluctuation, indicating that the availability of training sets was no longer the major limiting factor of accuracy.

Fig. 11 compares the PED prediction performance for different models under 2-, 4-, 6-, 8-, and 10-month data available conditions. It is pleasant to observe in Fig. 10 an obviously comparison between non-transfer model and transfer learning models. When only the 2- or 4-month dataset was available, the transfer-learning method rapidly improved prediction performance. Therefore, our proposed transfer learning approach has good potential for building load prediction tasks with limited data. When the 8-month dataset was available, the prediction error of the standalone model was significantly reduced. And with the data increasing, the advantage of transfer learning will be compromised.

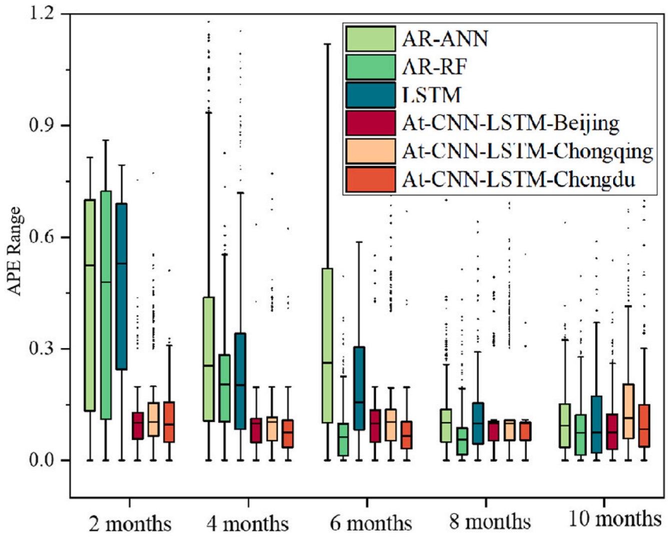


Fig. 11. PED prediction performance (APE) of four models.

The PED for the entire year predicted under 2- and 4-months data available conditions for different models are shown in Fig. 12. The blue curve represents the measured data, and the remaining colors show the predicted values for AR-ANN, AR-RF, LSTM, and At–CNN–LSTM. Because the 2-month dataset did not consider HVAC energy consumption, the predictive curves of AR-ANN, AR-RF, and LSTM were stable. Fig. 11 shows that the proposed models are more effective than the transfer the feature map from other models, especially in cooling reason.

Fig. 13 compares the TEC prediction performance for different models under 2-, 4-, 6-, 8-, and 10-month data available conditions. Similar to PED prediction, the proposed transfer learning model outperformed standalone models. However, when the data scale was 8 months or higher, transfer learning showed only marginal superiority, as the prediction error of the standalone model was small.

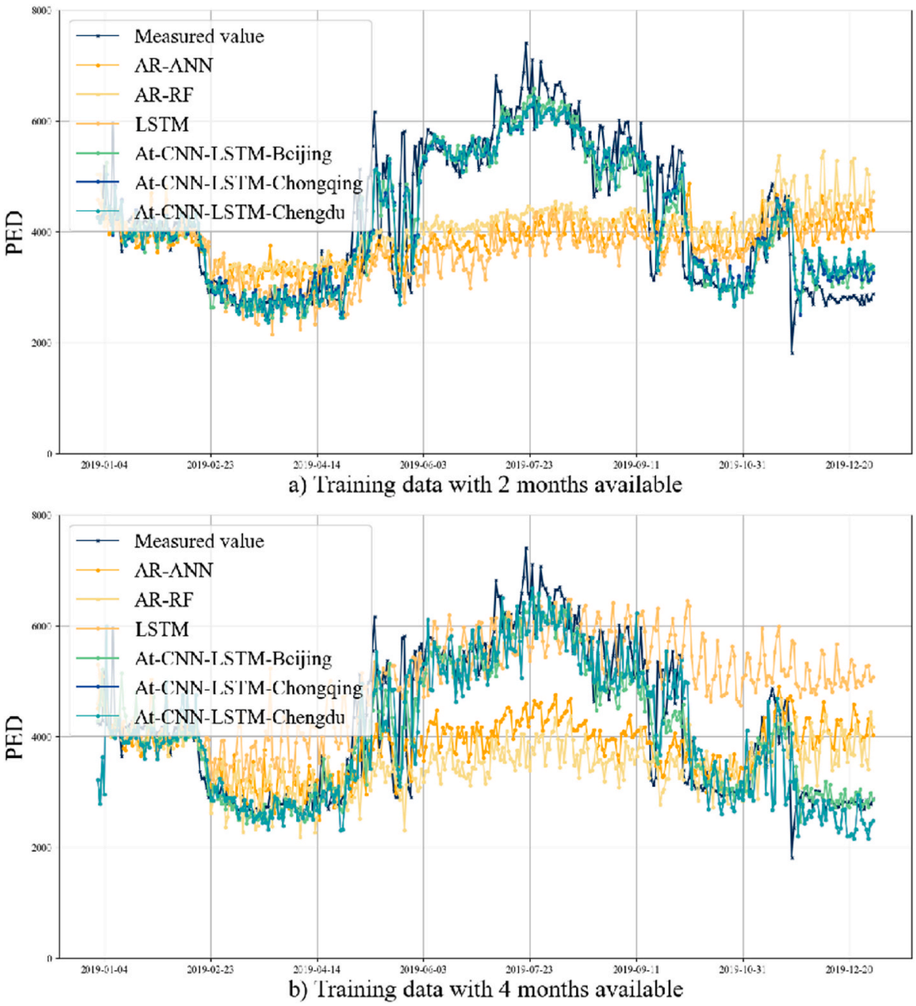
The TEC prediction performance for the entire year with 2- or 4-months of available data is shown in Fig. 14. The whole-year TEC trend of Harbin revealed that the abnormal curve of measured data in December was the major cause of the low prediction accuracy.

4. Conclusion and discussion

This study demonstrated that the CNN-LSTM model integrated with the attention mechanism ability could enhance the prediction accuracy and generalizability of building energy consumption models, which is especially important when the training data of the target building are insufficient.

The major findings and conclusions of this study are listed as follows.

1) When the training set included 12 months of available data, machine learning was a powerful tool for building energy prediction; the day-ahead prediction accuracy (CV-RMSE) was between 6.49% and 19.36%. Among the four models we compared, the proposed transfer learning approach (Attention–CNN–LSTM) proposed in this paper achieved the highest accuracy for all the prediction tasks. For the day-ahead prediction task, the CV-RMSE of the attention–CNN–LSTM was 6.49%, which was 2.90% lower than that of the second-performing model (LSTM). Attention–CNN–LSTM reduced CV-RMSE by 2.42% and 3.27%, respectively, compared with the standalone method for PED and TEC prediction tasks.
2) For PED and TEC predictions, the smallest CV-RMSE values of the proposed model were 8.44% and 7.21%, respectively, and the largest CV-RMSE differences were 8.73% and 9.61%, respectively. The results showed that the proposed transfer-learning approach effectively addressed the problem of LSTM being insensitive to long-distance historical sample information. However, the difference in



a) Training data with 2 months available

b) Training data with 4 months available

Fig. 12. PED prediction performance using 2 or 4 months of available data. a) Results for 2- month data. b) Results for 4-month data.

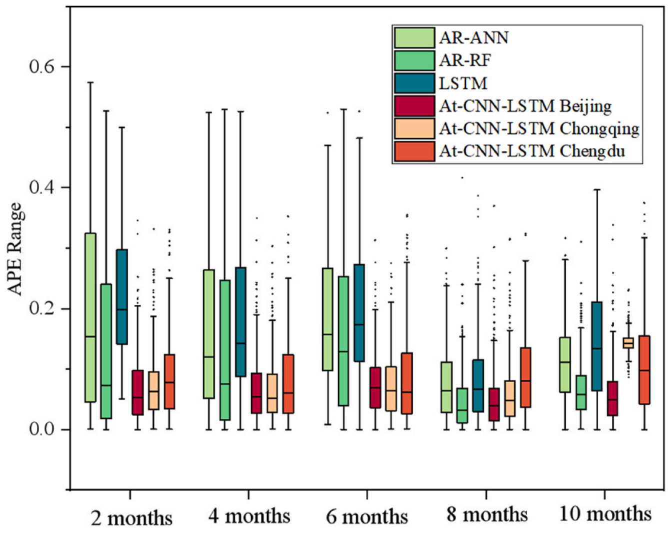
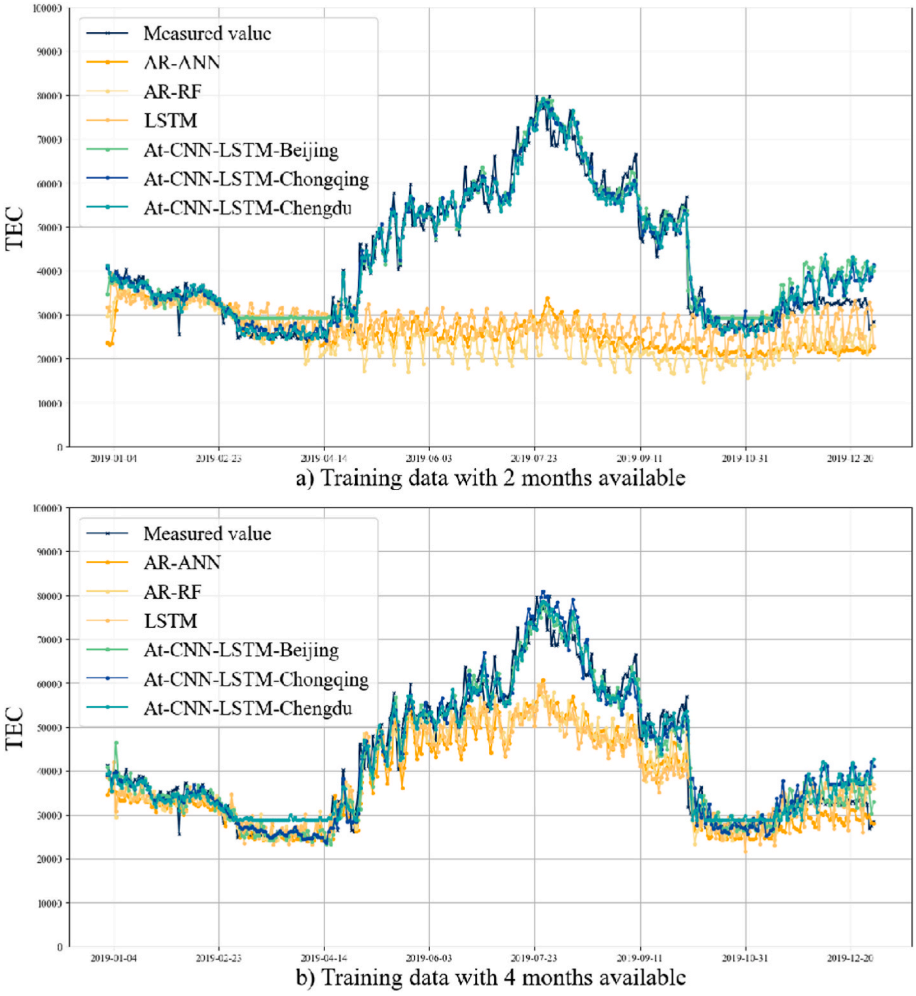


Fig. 13. TEC prediction performance (APE) of four models.

performance between the four algorithms was insignificant when the training data were sufficient.

3) When the training data were insufficient, the transfer learning methods significantly improved the performance compared with the standalone methods. Our proposed model's minimum prediction error (CV-RMSE) is 12.48% in PED prediction tasks and 10.78% in TEC prediction tasks with 2-month data available. The corresponding minimum prediction error (CV-RMSE) is 14.33% in PED prediction tasks and 11.32% in TEC prediction tasks with 4-month data available.
4) The effectiveness of the transfer learning method of the source domain selected by the MMD algorithm to the target domain was verified, thus facilitating the successful model transfer. Although the data availability cannot cover the cooling season, the attention mechanism-based CNN module is a suitable tool to extract the feature map from the source domain buildings, effectively improving the accuracy of the annual energy consumption predictive performance for target buildings. However, one shortcoming of this study is that the data scale used is limited. If the climate conditions of the source building are more similar to those of the target building, higher accuracy of the transfer learning model can be expected.

The limitation of this proposed model is that we need a large amount of data to establish the feature map set to migrate the buildings with insufficient data. In addition, only the impact of environmental factors has been considered in energy consumption forecasting. Still, the energy consumption of buildings is influenced by many variables, such as the time of use of pricing and building scales. Taking these variables into account when establishing prediction models will undoubtedly improve the accuracy of the entire model. This part of the content will be added in future research. Furthermore, a techno-economic analysis will be



a) Training data with 2 months available

b) Training data with 4 months available

Fig. 14. TEC prediction performance using 2 or 4 months of available data. a) Results for 2-month data. b) Results for 4-month data.

discussed in future work to address the application value of energy forecasting.

CRediT author statement

Yue Yuan: Methodology, Writing – original draft. **Zhihua Chen:** Validation. **Zhe Wang:** Writing – review & editing. **Yifu Sun:** Investigation, Data curation. **Yixing Chen:** Conceptualization, Supervision.

Declaration of competing interest

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (NSFC) through Grant (No. 51908204), Natural Science Foundation of Hunan Province of China through Grant (No. 2020JJ3008), and Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083).

References

[1] Deng Z, Chen Y, Yang J, Chen Z. Archetype identification and urban building energy modeling for city-scale buildings based on GIS datasets. Build Simulat 2022; 1547–59. https://doi.org/10.1007/s12273-021-0878-4.

[2] Li K, Ma M, Xiang X, Feng W, Ma Z, Cai W, et al. Carbon reduction in commercial building operations: a provincial retrospection in China. Appl Energy 2022;306: 118098. https://doi.org/10.1016/j.apenergy.2021.118098.

[3] Chen Z, Chen Y, Yang C. Impacts of large chilled water temperature difference on thermal comfort, equipment sizes, and energy saving potential. J Build Eng 2022; 49:104069. https://doi.org/10.1016/j.jobe.2022.104069.

[4] Chen Y, Yang C, Pan X, Yan D. Design and operation optimization of multi-chiller plants based on energy performance simulation. Energy Build 2020;222:110100. https://doi.org/10.1016/j.enbuild.2020.110100.

[5] Chen Y, Raphael B, Sekhar SC. Experimental and simulated energy performance of a personalized ventilation system with individual airflow control in a hot and humid climate. Build Environ 2016;96:283–92. https://doi.org/10.1016/j.buildenv.2015.11.036.

[6] Chen Y, Raphael B, Sekhar SC. Individual control of a personalized ventilation system integrated with an ambient mixing ventilation system. HVAC R Res 2012; 18:1136–52. https://doi.org/10.1080/10789669.2012.710059.

[7] Chellaswamy C, Ganesh Babu R, Vanathi A. A framework for building energy management system with residence mounted photovoltaic. Build Simulat 2021;14: 1031–46. https://doi.org/10.1007/s12273-020-0735-x.

[8] Chari A, Christodoulou S. Building energy performance prediction using neural networks. Energy Effic 2017;10:1315–27. https://doi.org/10.1007/s12053-017-9524-5.

[9] Lu C, Li S, Lu Z. Building energy prediction using artificial neural networks: a literature survey. Energy Build 2021:111718. https://doi.org/10.1016/j.enbuild.2021.111718.

[10] Fan C, Chen M, Tang R, Wang J. A novel deep generative modeling-based data augmentation strategy for improving short-term building energy predictions. Build Simulat 2022;15:197–211. https://doi.org/10.1007/s12273-021-0807-6.

[11] Ayele GT, Mabrouk MT, Haurant P, Laumert B, Lacarrière B. Optimal heat and electric power flows in the presence of intermittent renewable source, heat storage

and variable grid electricity tariff. Energy Convers Manag 2021;243. https://doi.org/10.1016/j.enconman.2021.114430.

[12] Sen P, Roy M, Pal P. Application of ARIMA for forecasting energy consumption and GHG emission: a case study of an Indian pig iron manufacturing organization. Energy 2016;116:1031–8. https://doi.org/10.1016/j.energy.2016.10.068.

[13] Jumin E, Basaruddin FB, Yusoff YBM, Latif SD, Ahmed AN. Solar radiation prediction using boosted decision tree regression model: a case study in Malaysia. Environ Sci Pollut Res 2021;28:26571–83. https://doi.org/10.1007/s11356-021-12435-6.

[14] Namazkhan M, Albers C, Steg L. A decision tree method for explaining household gas consumption: the role of building characteristics, socio-demographic variables, psychological factors and household behaviour. Renew Sustain Energy Rev 2020; 119:109542. https://doi.org/10.1016/j.rser.2019.109542.

[15] Yang YL, Che JX, Li YY, Zhao YJ, Zhu SL. An incremental electric load forecasting model based on support vector regression. Energy 2016;113:796–808. https://doi.org/10.1016/j.energy.2016.07.092.

[16] Pinanggih DH, Abdullah AG, Hakim DL. Prediction of energy consumption using artificial neural network method in one of shopping center in Cirebon city. IOP Conf Ser Mater Sci Eng 2021;1098:042011. https://doi.org/10.1088/1757-899x/1098/4/042011.

[17] Liu Y, Chen H, Zhang L, Feng Z. Enhancing building energy efficiency using a random forest model: a hybrid prediction approach. Energy Rep 2021;7:5003–12. https://doi.org/10.1016/j.egyr.2021.07.135.

[18] Zhou Y, Guo Q, Sun H, Yu Z, Wu J, Hao L. A novel data-driven approach for transient stability prediction of power systems considering the operational variability. Int J Electr Power Energy Syst 2019;107:379–94. https://doi.org/10.1016/j.ijepes.2018.11.031.

[19] Zhou C, Fang Z, Xu X, Zhang X, Ding Y, Jiang X, et al. Using long short-term memory networks to predict energy consumption of air-conditioning systems. Sustain Cities Soc 2020;55:102000. https://doi.org/10.1016/j.scs.2019.102000.

[20] Wen Q, Liu G, Rao Z, Liao S. Applications, evaluations and supportive strategies of distributed energy systems: a review. Energy Build 2020;225:110314. https://doi.org/10.1016/j.enbuild.2020.110314.

[21] Kamel E, Sheikh S, Huang X. Data-driven predictive models for residential building energy use based on the segregation of heating and cooling days. Energy 2020;206: 118045. https://doi.org/10.1016/j.energy.2020.118045.

[22] Cai L, Gu J, Jin Z. Two-layer transfer-learning-based architecture for short-term load forecasting. IEEE Trans Ind Inf 2020;16:1722–32. https://doi.org/10.1109/TII.2019.2924326.

[23] Pachauri N, Ahn CW. Weighted aggregated ensemble model for energy demand management of buildings. Energy 2023;263:125853. https://doi.org/10.1016/j.energy.2022.125853.

[24] Tian C, Ye Y, Lou Y, Zuo W, Zhang G, Li C. Daily power demand prediction for buildings at a large scale using a hybrid of physics-based model and generative adversarial network. Build Simulat 2022;15:1685–701. https://doi.org/10.1007/s12273-022-0887-y.

[25] Feng Y, Yao J, Li Z, Zheng R. Uncertainty prediction of energy consumption in buildings under stochastic shading adjustment. Energy 2022;254:124145. https://doi.org/10.1016/j.energy.2022.124145.

[26] Pinto G, Wang Z, Roy A, Hong T, Capozzoli A. Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives. Adv Appl Energy 2022;5:100084. https://doi.org/10.1016/j.adapen.2022.100084.

[27] Panigrahi S, Nanda A, Swarnkar T. A survey on transfer learning. Smart Innov Syst Technol 2021;194:781–9. https://doi.org/10.1007/978-981-15-5971-6_83.

[28] Fan C, Yan D, Xiao F, Li A, An J, Kang X. Advanced data analytics for enhancing building performances: from data-driven to big data-driven approaches. Build Simulat 2021;14:3–24. https://doi.org/10.1007/s12273-020-0723-1.

[29] Houidi S, Fourer D, Auger F, Sethom HBA, Miègeville L. Comparative evaluation of non-intrusive load monitoring methods using relevant features and transfer learning. Energies 2021;14:1–28. https://doi.org/10.3390/en14092726.

[30] Liu J, Zhang Q, Li X, Li G, Liu Z, Xie Y, et al. Transfer learning-based strategies for fault diagnosis in building energy systems. Energy Build 2021;250:111256. https://doi.org/10.1016/j.enbuild.2021.111256.

[31] Gao Y, Ruan Y, Fang C, Yin S. Deep learning and transfer learning models of energy consumption forecasting for a building with poor information data. Energy Build 2020;223:110156. https://doi.org/10.1016/j.enbuild.2020.110156.

[32] Miller C, Meggers F. The Building Data Genome Project : on an public data set from the Building non-residential Data Genome Project : open , public building an electrical meters data set from Assessing the feasibility of using electrical. Energy Proc 2017;122:439–44. https://doi.org/10.1016/j.egypro.2017.07.400.

[33] Lu Y, Tian Z, Zhou R, Liu W. A general transfer learning-based framework for thermal load prediction in regional energy system. Energy 2021;217:119322. https://doi.org/10.1016/j.energy.2020.119322.

[34] Ribeiro M, Grolinger K, ElYamany HF, Higashino WA, Capretz MAM. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. Energy Build 2018;165:352–63. https://doi.org/10.1016/j.enbuild.2018.01.034.

[35] Pb B, Bhuiyan MA, Zhang K, Song A. Transfer learning for leisure centre energy consumption prediction. Springer International Publishing; 2019. p. 112–23. https://doi.org/10.1007/978-3-030-22734-0.

[36] Tian Y, Sehovac L, Grolinger K. Similarity-based chained transfer learning for energy forecasting with big data. IEEE Access 2019;7:139895–908. https://doi.org/10.1109/ACCESS.2019.2943752.

[37] Grubinger T, Chasparis GC, Natschläger T. Online transfer learning for climate control in residential buildings. Eur Control Conf ECC 2016. https://doi.org/10.1109/ECC.2016.7810450. 2016 2017:1183–8.

[38] Fan C, Sun Y, Xiao F, Ma J, Lee D, Wang J, et al. Statistical investigations of transfer learning-based methodology for short-term building energy predictions. Appl Energy 2020;262:114499. https://doi.org/10.1016/j.apenergy.2020.114499.

[39] Fang X, Gong G, Li G, Chun L, Li W, Peng P. A hybrid deep transfer learning strategy for short term cross-building energy prediction. Energy 2021;215:119208. https://doi.org/10.1016/j.energy.2020.119208.

[40] Zhou D, Ma S, Hao J, Han D, Huang D, Yan S, et al. An electricity load forecasting model for Integrated Energy System based on BiGAN and transfer learning. Energy Rep 2020;6:3446–61. https://doi.org/10.1016/j.egyr.2020.12.010.

[41] Chen S, Ge H, Li H, Sun Y, Qian X. Hierarchical deep convolution neural networks based on transfer learning for transformer rectifier unit fault diagnosis. Meas J Int Meas Confed 2021;167:108257. https://doi.org/10.1016/j.measurement.2020.108257.

[42] Chinese National Standard. Standard of climatic regionalization for architecture (GB 50178-1993). 1993.

[43] Yang L, Lyu K, Li H, Liu Y. Building climate zoning in China using supervised classification-based machine learning. Build Environ 2020;171:106663. https://doi.org/10.1016/j.buildenv.2020.106663.

[44] Wang J, Li G, Chen H, Liu J, Guo Y, Sun S, et al. Energy consumption prediction for water-source heat pump system using pattern recognition-based algorithms. Appl Therm Eng 2018;136:755–66. https://doi.org/10.1016/j.applthermaleng.2018.03.009.

[45] Tian W, Zhu C, Sun Y, Li Z, Yin B. Energy characteristics of urban buildings: assessment by machine learning. Build Simulat 2021;14:179–93. https://doi.org/10.1007/s12273-020-0608-3.

[46] Babura BI, Adam MB, Rahim A. Analysis and Assessment of Boxplot Characters for Extreme Data Analysis and Assessment of Boxplot Characters for Extreme Data n.d. https://doi.org/10.1088/1742-6596/1132/1/012078.

[47] Amarasinghe K, Marino DL, Manic M. Deep neural networks for energy load forecasting. IEEE Int Symp Ind Electron 2017:1483. https://doi.org/10.1109/ISIE.2017.8001465. –8.

[48] Wang L, Lee EWM, Yuen RKK. Novel dynamic forecasting model for building cooling loads combining an artificial neural network and an ensemble approach. Appl Energy 2018;228:1740–53. https://doi.org/10.1016/j.apenergy.2018.07.085.

[49] Wang Z, Wang Y, Zeng R, Srinivasan RS, Ahrentzen S. Random Forest based hourly building energy prediction. Energy Build 2018;171:11–25. https://doi.org/10.1016/j.enbuild.2018.04.008.

[50] Zekić-Sušac M, Has A, Knežević M. Predicting energy cost of public buildings by artificial neural networks, CART, and random forest. Neurocomputing 2021;439: 223–33. https://doi.org/10.1016/j.neucom.2020.01.124.

[51] Wang JQ, Du Y, Wang J. LSTM based long-term energy consumption prediction with periodicity. Energy 2020;197. https://doi.org/10.1016/j.energy.2020.117197.

[52] Thang Luong, Hieu Pham, Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. Lisbon, Portugal: Association for Computational Linguistics; 2015. p. 1412–21.

[53] Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

[54] ASHRAE. Guideline 14-2014, measurement of energy and demand savings, vol. 4; 2014.