https://doi.org/10.1016/j.apenergy.2020.114499
This is the Pre-Published Version.

Statistical investigations of transfer learning-based methodology for short-term building energy predictions

Cheng Fan[1], Yongjun Sun[2], Fu Xiao[3,*], Jie Ma[4], Dasheng Lee[5], Jiayuan Wang[1], Yen Chieh Tseng[5]

[1]Sino-Australia Joint Research Center in BIM and Smart Construction, Shenzhen University, Shenzhen, China

[2]Division of Building Science and Technology, City University of Hong Kong, Hong Kong, China

[3]Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[4]School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China

[5]Department of Energy and Refrigerating Air-conditioning Engineering, National Taipei University of Technology, Taiwan, China

[*] E-mail: linda.xiao@polyu.edu.hk

Abstract

The wide availability of massive building operational data has motivated the development of advanced data-driven methods for building energy predictions. Existing data-driven prediction methods are typically customized for individual buildings and their performance are highly influenced by the training data amount and quality. In practice, buildings may only possess limited measurements due to the lack of advanced monitoring systems or data accumulation time. As a result, existing data-driven approaches may not present sufficient values for practical applications. A novel solution can be developed based on transfer learning, which utilizes the knowledge learnt from well-measured buildings to facilitate prediction tasks in other buildings. However, the potentials of transfer learning-based methods for building energy predictions have not been systematically examined. To address this research gap, a transfer learning-based methodology is proposed for 24-hour ahead building energy demand predictions. Experiments have been designed to investigate the potentials of transfer learning in different scenarios with different implementation strategies. Statistical assessments have been performed to validate the value of transfer learning in short-term building energy predictions. Compared with standalone models, the transfer learning-based methodology could reduce approximately 15% to 78% of prediction errors. The research outcomes are useful for developing advanced transfer

© 2020. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/.

learning-based methods for typical tasks in building energy management. The insights obtained can help the building industry to fully realize the value of existing building data resources and advanced data analytics.
Keywords: Building energy predictions; Transfer learning; Deep learning; Data-driven models; Smart building energy management.

1. Introduction

The wide availability of massive building operational data has motivated the development of advanced data-driven methods for smart building energy management. As the fundamental basis for many building energy management tasks, short-term building energy predictions have drawn great attentions from both building professionals and academics [1, 2]. A large number of studies, which utilized advanced machine learning techniques, have been performed to validate the usefulness of data-driven building energy prediction methods [3, 4]. Such methods have gained increasing popularities due to their implementation flexibilities and excellence in describing complex relationships among data variables [5, 6].

However, the practical value of advanced data-driven methods for building energy predictions can be greatly limited due to two practical constraints. Firstly, the prediction performance is highly influenced by the training data amount and the intrinsic complexity of machine learning algorithms. For instance, using a complicated deep neural network typically results in more accurate predictions than linear regressions. Nevertheless, the former also requires more training data to avoid the essential problem in predictive modeling, i.e., over-fitting [7]. Such high data dependence can impose great challenges in practice, especially for buildings with limited measurements. Secondly, existing studies are typically customized for a specific building. As a result, the model developed cannot be applicable for other buildings with different operating characteristics.

Transfer learning, which aims to improve the performance of a target task by applying the knowledge learnt from other tasks, can be adopted to tackle these challenges [8, 9]. It has been successful used in several fields, such as image classification [10], text sentiment analysis [11], object detection [12] and economic analysis [13]. Transfer learning is motivated by the fact that people can intelligently apply knowledge learnt from one domain to tackle challenges in other domains [9]. Rather than developing a standalone model for each task, it utilizes the knowledge learnt from previous tasks to a new one and thereby, reducing the need for training data and computation resources [14]. In essence, transfer learning helps to break the fundamental assumption in conventional machine learning, i.e., the training and testing data

should be drawn from the same feature space and have similar data characteristics. It is extremely useful when there are limited training data, or the data collection process is expensive and time-consuming.

With building operational data resources becoming increasingly prevalent, a natural question arises as whether transfer learning can be applied to facilitate data-driven tasks in building energy management. In terms of building energy predictions, it is a very attractive idea to utilize existing data resources from well-measured buildings to facilitate the reliable prediction model development for other buildings. Previous studies mainly focus on developing customized solutions for individual buildings [15, 16]. To the best of our knowledge, only few studies have been performed to address the potential of transfer learning in the building field. Grubinger et al. developed a generalized online transfer learning framework for improving the temperature predictions in residential buildings [17]. The research results validated the usefulness of transfer learning in utilize existing building operational data. Ribeiro et al. adopted transfer learning to perform cross-building energy forecasting [18]. The method was validated using a school building and a performance boost of 11.2% was reported. Mocanu et al. adopts reinforcement learning and transfer learning to predict building energy consumptions [19]. The method was proved to be useful in transferring knowledge learnt for identifying new building operating behaviors.

From the authors' perspectives, the transfer learning-based solutions for building energy predictions are especially promising for two practical scenarios. The first is for existing buildings with limited building operational data due to the lack of advanced building automation systems and regular data collection activities. In such a case, even though the building has existed for a fairly long time, the training data available for model development is limited due to large collection intervals (e.g., monthly or yearly high-level measurements), errors or malfunctions in sensors and data storage systems. The second is for new buildings which have only experienced a few operating conditions with limited data accumulation time. In such a case, the measurements at hand cannot fully describe seasonalities or building operating characteristics, making it infeasible to establish advanced data-driven models without a time-consuming data accumulation process.

To systematically reveal the usefulness of transfer learning in building energy predictions, this paper proposes a novel methodology for 24-hour ahead building energy predictions. Data experiments are designed considering different learning scenarios and transfer learning implementation strategies. The paper is organized as follows. Section 2 provides the theoretical

background. Section 3 describes the research outline and experiment setups. The results are shown and discussed in Section 4. Conclusions are drawn in Section 5.

2. Theoretical background

2.1 Principle of transfer learning

Transfer learning is motivated by the inherent human ability to utilize knowledge learnt from one task to others. As shown in Fig. 1, rather than learning from scratch for each new task, transfer learning can improve the overall learning efficiency by leveraging the knowledge learnt from related tasks. It is of great significance for practical applications as advanced learning algorithms typically require large amounts of training data for reliable development while it can be time-consuming, expensive or even impossible for additional data accumulation. As described in [9], the framework of transfer learning can be formally defined as follows. A domain D consists of two components as $D = \{X, P(X)\}$, where X and $P(X)$ represent features and their marginal probability respectively. A task T also consists of two components and is defined as $T = \{Y, P(Y|X)\}$, where Y represents the label and $P(Y|X)$ is the conditional probability of Y given X. Given a source domain D_s and a corresponding source task T_s, a target domain D_t and a target task T_t, transfer learning aims to learn the target conditional probability distribution $P(Y_t|X_t)$ in D_t with the help of knowledge learnt from D_s and T_s, where D_s and D_t, T_s and T_t are not identical.

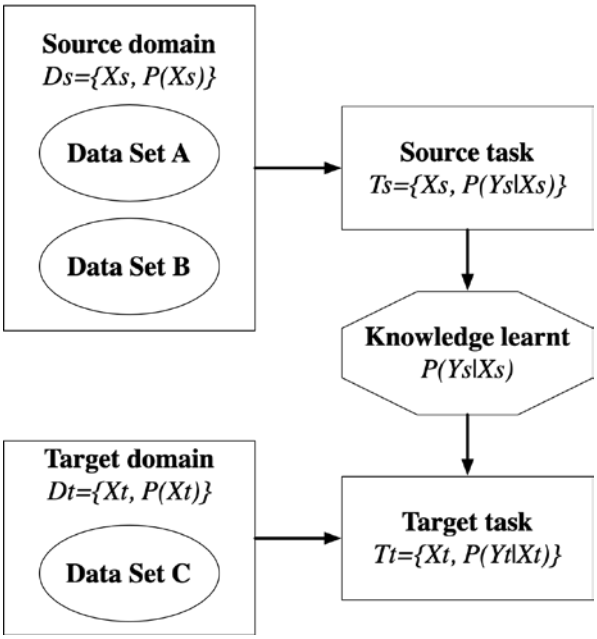

Fig. 1 The transfer learning principle

In general, there are three types of transfer learning based on the domains and tasks at hand, i.e., inductive, transductive and unsupervised transfer learning [9]. In inductive transfer

learning, both the source and target domains have labeled data, yet the source and target tasks are different. Taking buildings as examples, inductive transfer learning can be applied when the building variables collected in different buildings are the same, yet some buildings focus on predicting building cooling loads while the others are interested in predicting indoor thermal comforts.

In transductive transfer learning, the source and target tasks are the same, yet the source and target domains are different. In this setting, the source domain has sufficient labeled data while the target domain has none. Taking buildings as examples, transductive transfer learning can be applied to perform system fault detection in one building with few faulty measurements by utilizing faulty measurements in other buildings.

The settings of unsupervised transfer learning are similar to that in inductive learning, i.e., the source and target domains are the same with different but related tasks. However, there are no labeled data in both domains and the aim is to explore the intrinsic data characteristics in different domains. Taking buildings as examples, it can be used to facilitate the automatic identification of typical operation patterns through clustering analysis. In such a case, the vast amount of source data can produce more stable estimations of clustering centroids and boundaries, which can enhance the clustering reliability and quality in the target domain.

2.2 Typical methods for knowledge transfer

For each type of transfer learning, different methods can be applied to transfer knowledge between source and target domains. In general, there are four groups of methods [9]. The first is instance-based transfer. The rationale behind is that a certain subset of source domain data can be used for the target domain task. One common practice is to apply re-weighting and importance sampling techniques to incorporate source domain data into the target task training process [8]. It can be used in either inductive or transductive transfer learning. The second is feature-based transfer, which aims to minimize domain divergence by deriving good feature representations from the source to task domains. A popular approach is to identify a latent feature space from the source domain, based on which the marginal distributions between two domains are minimized [20]. Such method is compatible with all three transfer learning types introduced in Section 2.1. The third is relational knowledge-based transfer, which is specifically designed to transfer relational relationships in non-independent and identically distributed data, such as social network data and text data [21]. It has been mainly used in inductive transfer learning [9]. The fourth is parameter-based transfer, which assumes that models developed for relevant tasks would share similar model parameters or priori distributions of hyper-parameters [9, 22]. In such a case, the knowledge learnt from the source

task is transferred to another task as shared model weights. It can be applied in both inductive and transductive transfer learning.

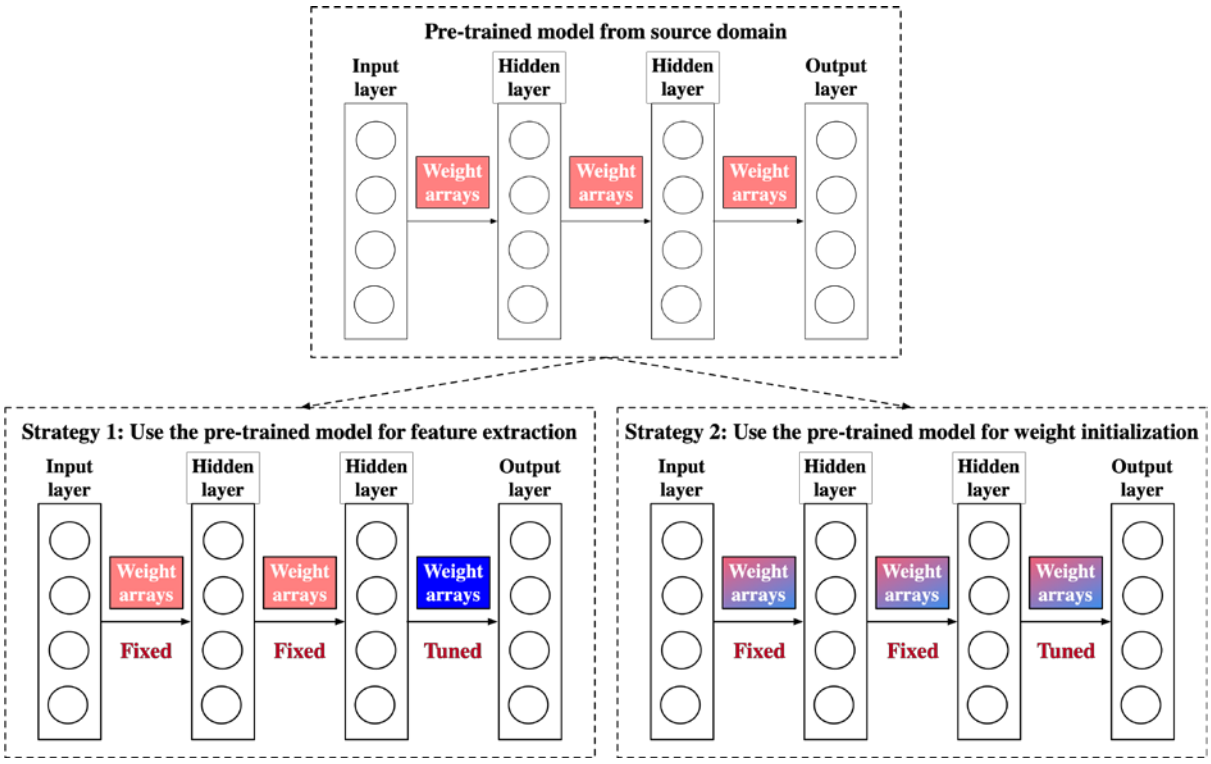

Fig. 2 Two common strategies for network-based transfer learning

The recent success of deep learning has prospered a new category of transfer learning, i.e., network-based transfer learning [22]. It is developed to solve the fundamental challenge of developing complicated deep learning models, i.e., insufficient training data. The network-based transfer learning belongs to the category of parameter-based transfer learning, as the main idea is to reuse the partial network trained in the source domain to facilitate the target task. As shown in Fig. 2, the network-based transfer can be implemented using two strategies [23]. The first is to utilize the pre-trained model developed in the source domain for feature extraction [24]. In such a case, all the model weights are fixed except for the output layer or the last few layers. New layers can be added for target domain adaption and their weights are learnt from target data. Since the number of weights to be learnt is greatly reduced, the development of new model requires much less training data and computation resources. The second is to utilize the pre-trained model for weight initialization [25]. In such a case, the weights of the pre-trained model are used for initialization only and can be adjusted through a fine-tuning process.

2.3 Data-driven methods for short-term building energy predictions

Data-driven methods have gained increasing popularities in building energy predictions due to their flexibility for practical applications [26]. The main idea is to adopt an inverse modeling

approach to describe the relationships between model inputs and outputs, i.e., model parameters are fitted based on actual building operational data rather than predefined by domain expertise.

The most widely used input data can be summarized into three types, i.e., time data, environmental data and historical data [27, 28]. Time data (e.g., *Day type* and *Hour*) are typically selected as proxies for indoor occupancy [29, 30]. Environmental data describe the indoor and outdoor conditions, such as the outdoor dry-bulb temperature and relative humidity. They are influential to building energy predictions due to their close relationships with the operating conditions of air-conditioning systems. The third is historical energy consumption data, which consist of actual measurements at previous time steps. A window-based method has been widely used for constructing input data set [26]. For instance, if a window size of w is specified, the building energy consumptions at time steps T-w to T-1 will be used as inputs to predict the energy consumption at time step T. The window size can be defined based on engineering experience or data-driven tests, such as autocorrelations and spectral density estimations [31, 32].

It is worth mentioning that the interactions between the above-mentioned input variables and building energy consumptions can be nonlinear and complicated. To enhance the prediction performance, advanced machine learning techniques have been utilized, such as fully connected neural networks [26], recurrent neural networks [6], support vector regression [15, 33] and ensemble trees [4]. The main challenge in developing reliable data-driven prediction models is to avoid overfitting, which typically results from training complicated models with insufficient data. Such problem can be partially alleviated by two approaches. The first is to use regularization techniques, such as the dropout technique for deep neural networks and the Lasso regularization for multiple linear regressions [7, 34]. The second is to elaborately design a model development scheme. One common approach is to divide the whole data into three parts, i.e., training, validation and testing data sets, and use cross validation for model parameter optimization. Another popular approach is to adopt data augmentation techniques to generate synthetic data for model training [35]. Rather than focusing on the data available in each individual building, transfer learning resorts to data in other buildings for help and therefore, providing a promising alternative to fully realize the value of existing data resources in the building field.

3. Research methodology

3.1 Research outline

This study proposes a transfer learning-based methodology for 24-hour ahead building energy demand predictions with the aim of addressing the following three essential questions:

(1) Can transfer learning benefit the energy prediction task for individual buildings by utilizing knowledge learnt from other buildings?

(2) Which type of the network-based transfer learning implementation strategies results in better performance for short-term building energy predictions, i.e., using the pre-trained model for feature extraction or weight initialization?

(3) Are there any guidelines to utilize transfer learning for short-term building energy predictions in different learning scenarios and data availabilities?

The research outline is depicted in Fig. 3. The prediction task is defined to predict 24-hour ahead building energy consumptions. Firstly, given data sets retrieved from different buildings, around 80% of the buildings are randomly selected as the source domain for developing a pre-trained model, while the other 20% are used as target domain for evaluating the usefulness of transfer learning. Model optimization is performed to determine the optimal model architecture. Data experiments are then conducted in the target domain to quantitatively evaluate the value of transfer learning in different learning scenarios and using different implementation strategies.

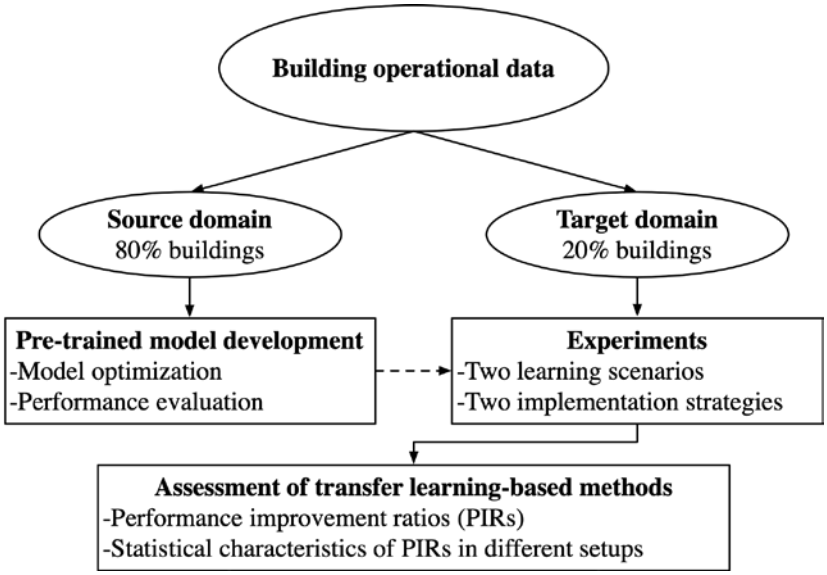

Fig. 3 The research outline

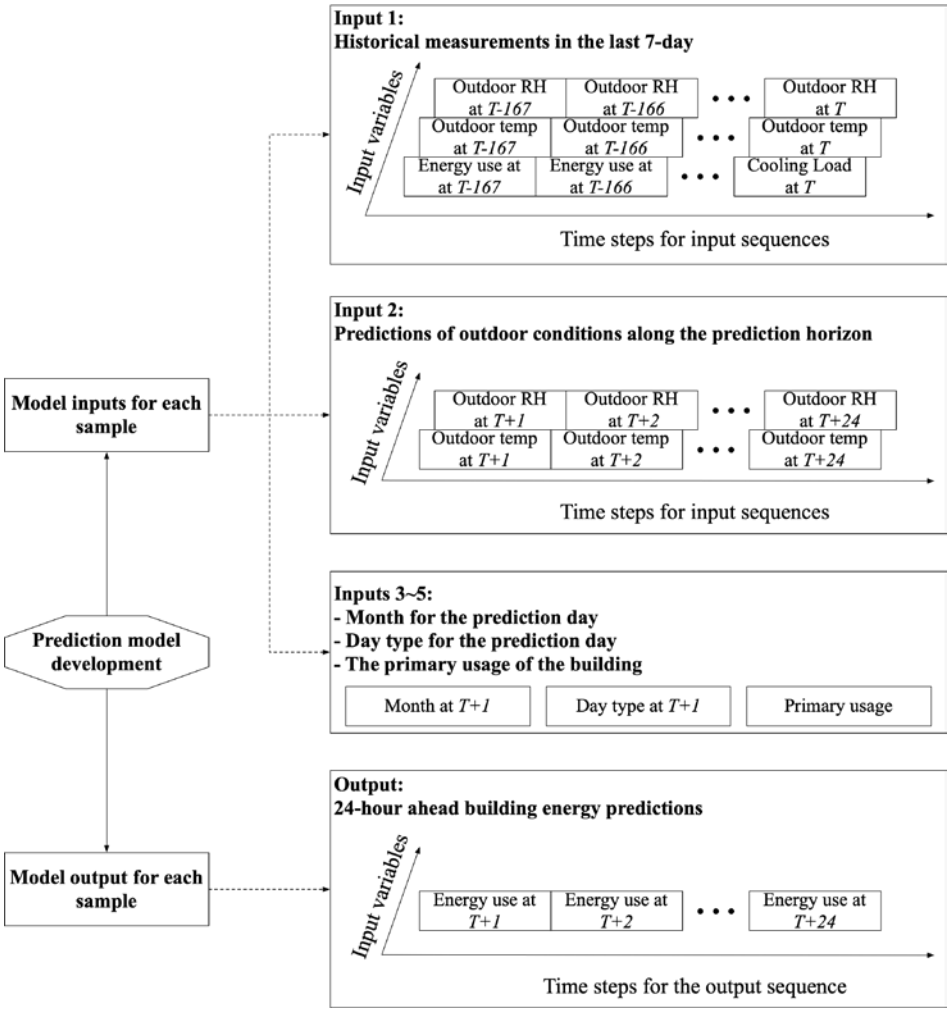

Fig. 4 Overview of model inputs and output

3.2 Pre-trained model development

The pre-trained model is developed based on the building operational data in the source domain. The prediction task is defined as the 24-hour ahead building power consumptions with one-hour interval. In other words, the model output for each sample consists of 24 values. The pre-trained model is developed with two considerations. Firstly, the input data should be easy to collect in practice while providing sufficient information to generate reliable and accurate predictions. Secondly, the model architecture should be capable of capturing high-level features and nonlinear temporal relationships for accurate time series predictions. The details are explained as follows.

3.2.1 Model inputs and outputs

As shown in Fig. 4, the input data used in this study can be categorized into five parts. Considering that the target task is in essence time series predictions, the first part includes the historical measurements of power consumptions and outdoor conditions. The outdoor dry-bulb temperature and relative humidity are selected to describe outdoor conditions as they are

closely related to building system operations and typically available in practice. Considering that buildings normally present significant weekly and daily patterns, the window size is set as 168 (i.e., 24×7). In other words, the historical measurements from the last seven days are utilized to predict the next-day building power consumptions. As a result, the first input consists of three time series sequences, each with a length of 168. The second part takes into account the outdoor conditions in the upcoming 24-hour and consists of two time series sequences (i.e., outdoor dry-bulb temperature and relative humidity) each with a length of 24. The third and fourth inputs represent *Month* (i.e., January to December) and *Day type* (i.e., Monday to Sunday) for the prediction day respectively. These two variables are treated as categorical variables with 12 and 7 levels respectively. They are used as indicators for seasonalities and indoor occupancy. The fifth input is a categorical variable specifying the building primary usage type. It can be used to partially represent occupant behaviors in different building types. The model output for each sample has 24 values, indicating the hourly building power consumptions for the upcoming 24-hour.

3.2.2 Model architecture

Given the model inputs and building operational data in the source domain, optimization is performed to determine the optimal network architecture for 24-hour ahead building power predictions. Considering the intrinsic data nature of inputs and outputs, the general model architecture consists of three main blocks. Firstly, one-dimensional (i.e., denoted as 1D) convolutional layers are utilized to automatically extract local temporal features from time series. It serves as a data preprocessing step and has been reported to be useful in reducing computational costs and enhancing model performance [35]. The number of 1D convolutional layers, the number of filters, kernel sizes, strides and activation function types are optimized through grid-search.

Secondly, recurrent layers are adopted to capture interactions among temporal features for accurate predictions. The recurrent units are Long Short-Term Memory (*LSTM*) units, considering its excellence in modeling long-term dependency and tackling the vanishing or exploding gradient problem [36]. The activation function type, the number of recurrent layers and recurrent units are optimized based on grid-search. In addition, to further enhance the performance of recurrent units, bidirectional operations and dropout techniques are integrated for model development. Bidirectional operations are used to reverse input temporal orders for incorporating both past and future information for accurate modeling [37]. Dropout is a popular regularization technique for tacking the overfitting problem in deep learning models. It refers to the process of randomly setting part of the model parameters as zeros during model training

[38]. Previous studies have shown that rather than applying different dropout masks at different time steps, a universal dropout mask across the input, output and recurrent layers are more suitable for regularizing recurrent models [39]. Both dropout operations are adopted in this study. The detailed grid-search settings for pre-trained model optimization are summarized in Table-1.

Thirdly, each of the three categorical input variables, i.e., *Month*, *Day Type* and *Primary Usage*, is transformed into numerical formats for model development. A conventional approach is one-hot encoding, where a matrix of *L-1* columns is generated for a categorical variable with *L* levels. Such approach is easy to implement yet may cause high data sparsity when the categorical variables have too many levels. In the field of deep learning, an embedding layer can be used to create dense representations for categorical variables [7, 35]. In this study, categorical inputs are transformed into numerical values through embedding layers. The embedding information is then integrated with recurrent layer outputs through element-wise multiplication.

Table-1 The grid-search settings for pre-trained model optimization

Parameters	Grid-search values
The number of 1D convolutional layers	0, 1, 2
The filter number in each 1D convolutional layer	100, 200, 300
The activation function in hidden layers	ReLU, Sigmoid, Tanh
Kernel size of 1D convolutions	4, 6, 8
Stride size of 1D convolutions	1, 2
The number of recurrent layers	1, 2
The number of recurrent units in each recurrent layer	24, 48, 72, 96
Bidirectional operations	Yes, No
Recurrent dropout	0.0, 0.1, 0.2, 0.3, 0.4, 0.5
Dropout	0.0, 0.1, 0.2, 0.3, 0.4, 0.5

3.3 Experiment setups

To comprehensively and quantitatively investigate the usefulness of transfer learning in short-term building energy predictions, experiments are performed for each building in two learning scenarios with different implementation strategies. The experiment outlines for each learning scenario are illustrated in Figs. 5 and 6 respectively, where the building operational data in various months are denoted as circle points with different colors. The details are explained in the following two subsections.

3.3.1 Setups for learning scenarios

Two learning scenarios are simulated for practical applications. The first simulates cases when the building has existed for a long time, yet the training data available is insufficient due to the lack of advanced monitoring systems or regular data collections. The second simulates cases for new buildings, where building operational data recorded cannot adequately describe building operating conditions or seasonalities. Different levels of data availabilities are specified for data experiments in each learning scenario.

As illustrated in Fig. 5, in *Learning Scenario A*, four cases are simulated by assuming 20%, 40%, 60% and 80% data availabilities for model training. For instance, if a building has m measurements in total, the first simulation case will randomly select $0.2m$ measurements across the whole year as the training data, while the remaining $0.8m$ will be used as testing data. Since the training data are selected through random sampling, the relative frequencies of measurements in different months should be similar across different data availabilities.

In *Learning Scenario B*, four cases are simulated by assuming 2-month, 5-month, 7-month and 10-month data availabilities for model training. The month numbers are specially designed to approximate the data proportions in *Learning Scenario A*. Taking Fig. 6 as an example, the first simulation case selects the building operational data from January and March as the training data, while the data in the other 10-month are used as testing data for performance evaluation.

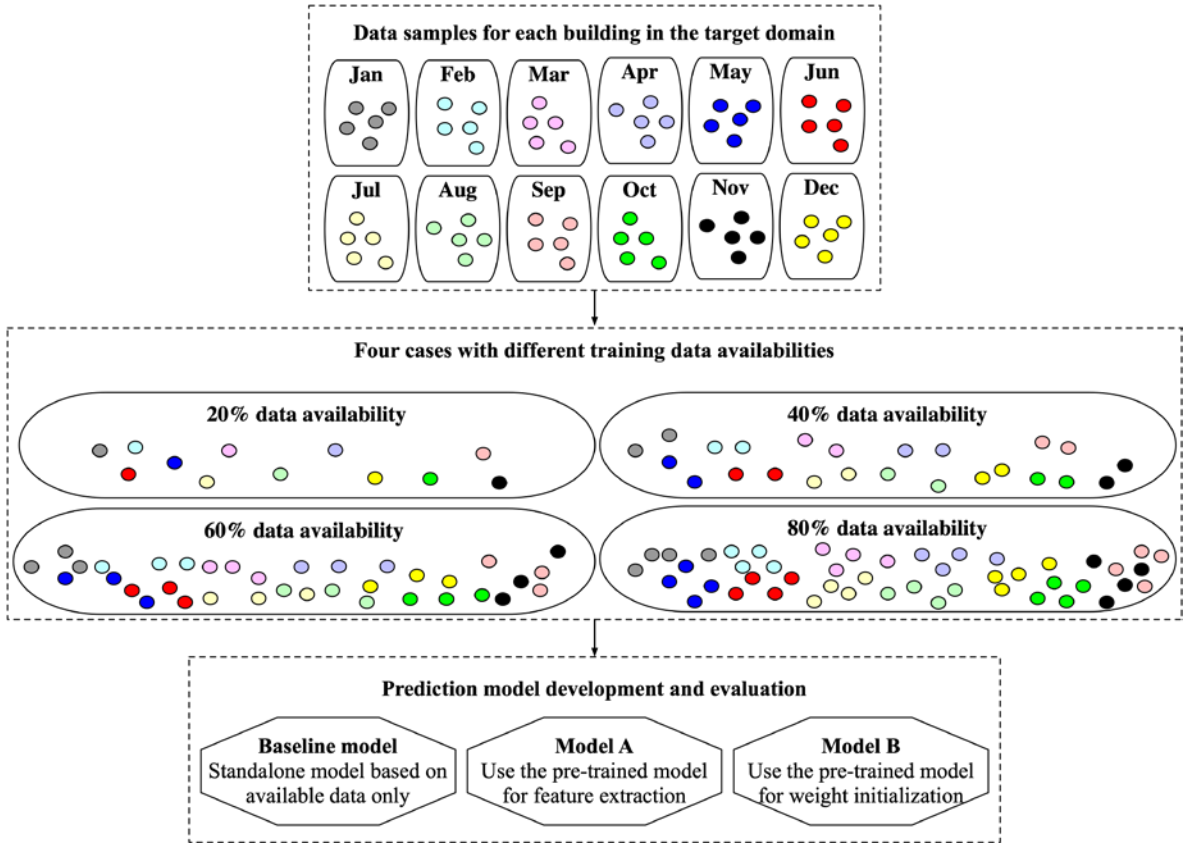

Fig. 5 Experiment setups for Learning Scenario A

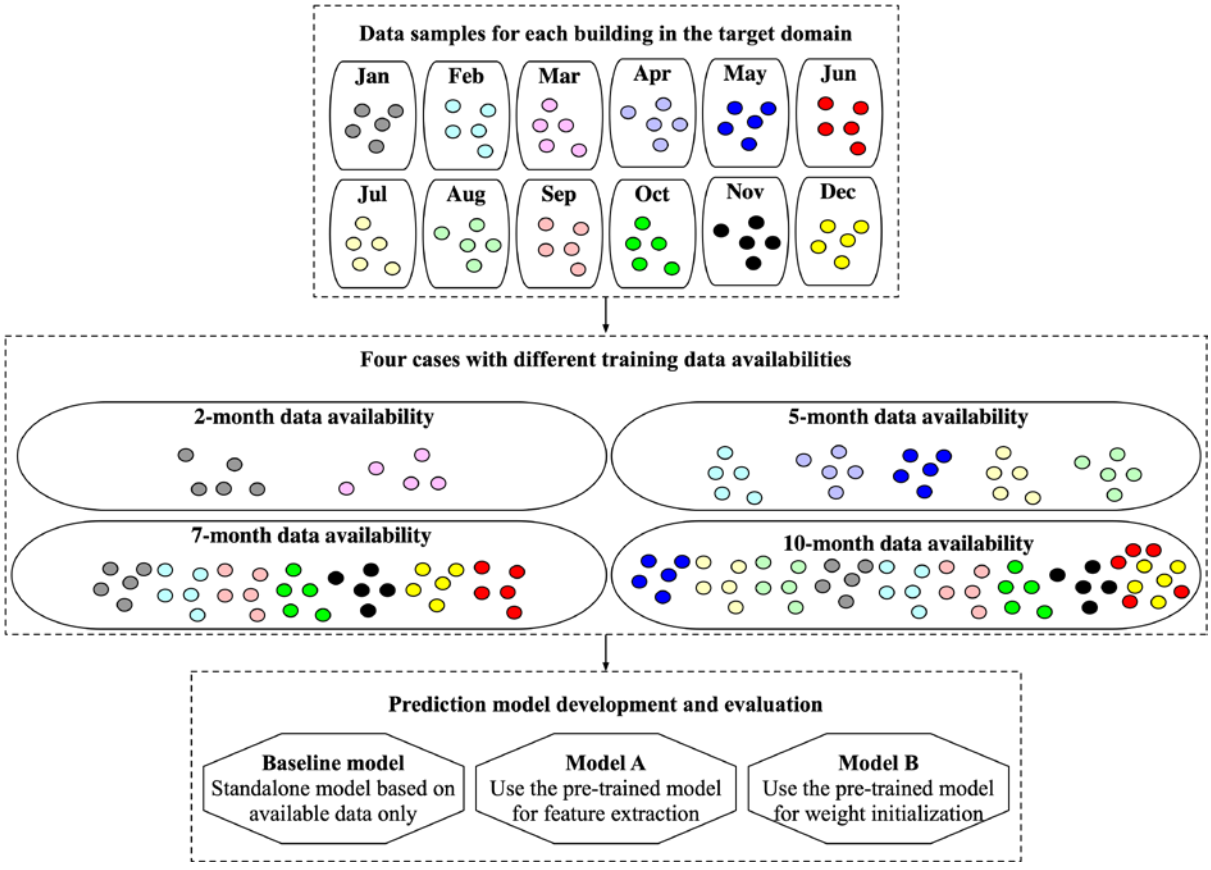

Fig. 6 Experiment setups for Learning Scenario B

3.3.2 Implementation strategies for knowledge transfer

The performance of two network-based implementation strategies for knowledge transfer is also investigated in this study. For each building in the target domain, a baseline model is established by directly developing a standalone model based on the training data available. The baseline model has the same architecture of the pre-trained model and the resulting prediction performance is used as performance benchmarks.

The first implementation strategy is to utilize the pre-trained model for feature extraction and the resulting models are denoted as *Model A*. In such a case, the whole pre-trained model except for the output layer are directly used with fixed parameters. It is connected with a new output layer with 24 neurons for predictions and the parameters in the last layer are learnt based on training data available. The second implementation strategy is to utilize the pre-trained model for weight initialization and the models developed are denoted as *Model B*. In such a case, the model weights of the pre-trained model are used to initialize the training process and will be fine-tuned using the training data available.

3.4 Performance evaluation

The root mean squared error (*RMSE*) and the coefficient of variation of the root mean squared error (*CV-RMSE*) are used to present prediction accuracy. These two metrics are calculated based on Eqs. 1 and 2 respectively, where $\hat{y}$ and y are predicted and actual values, n is the total sample size. In addition, a novel metric, i.e., performance improvement ratio (*PIR*), is defined to quantify the usefulness of transfer learning under different experiment settings. *PIR* is calculated based on Eq. 3, where $RMSE_1$ is the root mean squared error obtained by the standalone model and $RMSE_2$ is the root mean squared error obtained by utilizing the pre-trained model. If the knowledge transferred are beneficial for building energy predictions, $RMSE_2$ should be smaller than $RMSE_1$ and therefore, *PIR* should be positive. The value of knowledge transferred in the target task is positively correlated with the *PIR*. A negative *PIR* indicates that transfer learning does not lead to any performance enhancement.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

$$CV - RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \Big/ \frac{\sum_{i=1}^{n} y_i}{n} \quad (2)$$

$$PIR = \frac{RMSE_1 - RMSE_2}{RMSE_1} \quad (3)$$

4. Results and discussions

4.1 Data description and preparation

The building data set from the Building Data Genome Project was adopted for analyses [40]. The data set consists of 507 non-residential buildings. For each building, there are three types of data, i.e., building meta data, building power consumptions and outdoor conditions. The building meta data provides the general building information, such as the building location, primary usage type, the occupancy number, the total floor area, locations and etc. As shown in Fig. 7, the buildings are mainly located in America and Europe. New York and London have the largest building numbers (i.e., 151 and 143 respectively), while only 5 buildings in Asia are included. The buildings are categorized into five primary usage types and their frequencies are shown in Fig. 8. In total, there are 156 Offices, 105 primary or secondary school classrooms, 95 university laboratories, 81 university classrooms and 70 university dormitories. A heatmap regarding to the building primary usage types and locations is presented as Fig. 9. For instance, the data set includes 74 primary and secondary schools in London and 42 office buildings in Chicago. Each building has one-year measurements of hourly power consumptions. Since there are evident differences in building scales and occupancy numbers, the power consumptions also present large variations among buildings. As an example, Fig. 10 illustrates the average daily power patterns of 81 university classroom buildings on Monday. Each grey dashed line presents the average Monday pattern of a building and the red solid line represents the overall mean values at each time step. The outdoor conditions are described with a ten-minute time interval using various meteorological variables, such as the dry-bulb temperature, relative humidity, outdoor pressure and wind speed.

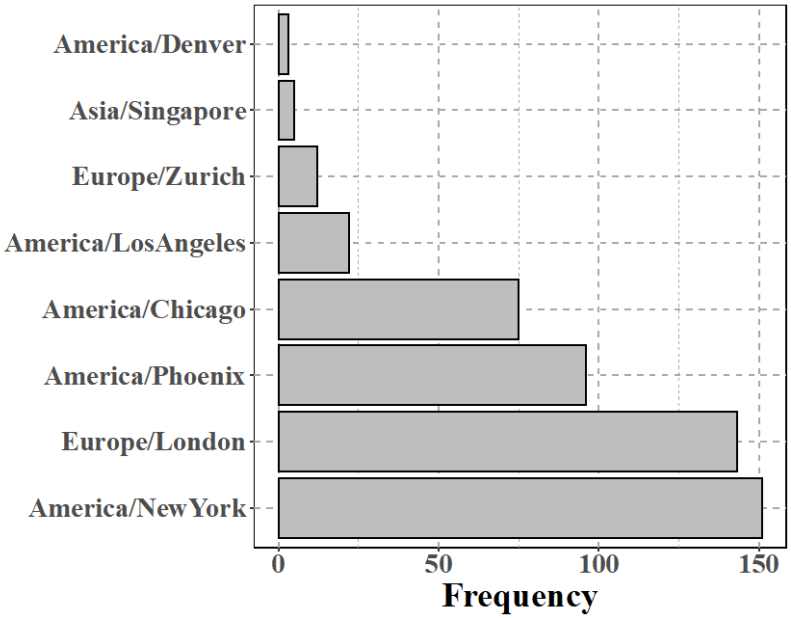

Fig. 7 The building numbers in different locations

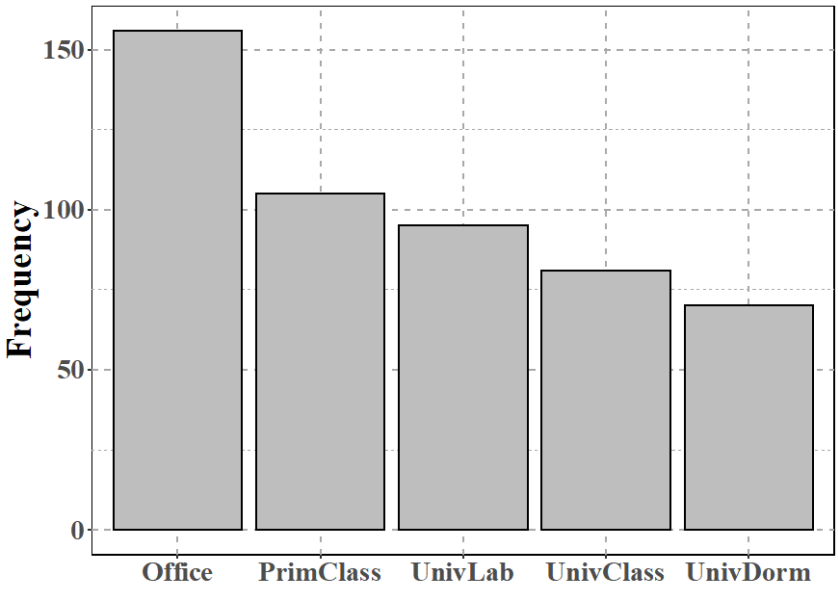

Fig. 8 The building numbers of different primary usage types

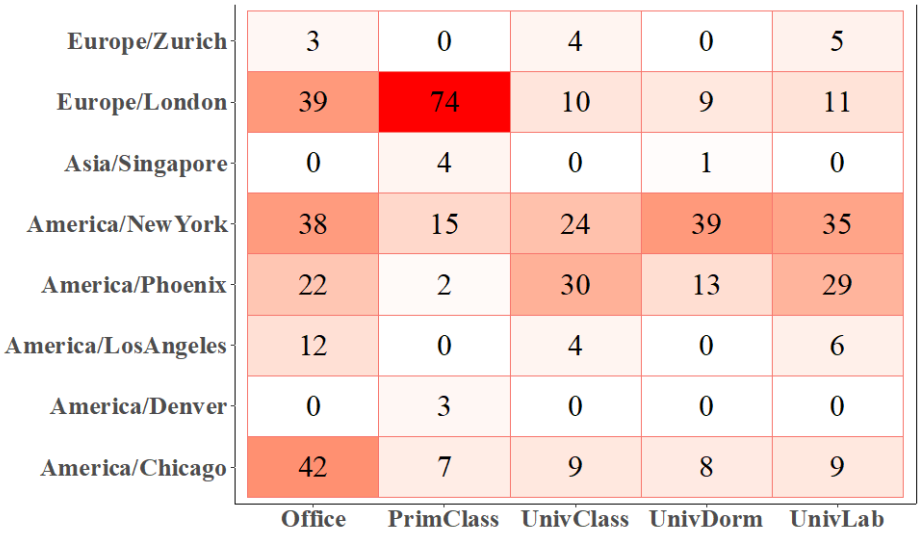

Fig. 9 The building numbers of different primary usage types and locations

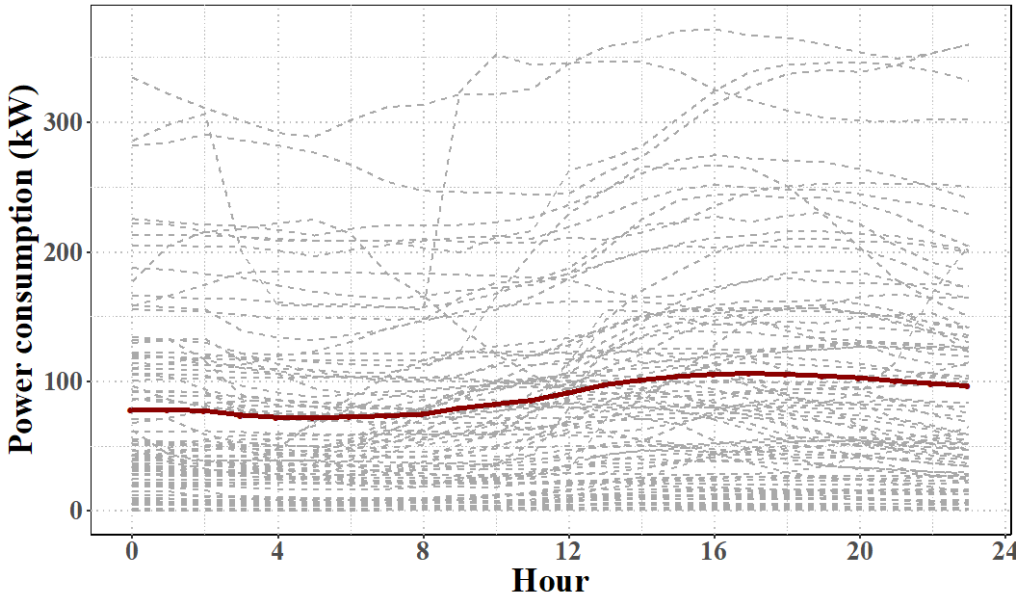

Fig. 10 Average power patterns of university classroom buildings on Monday

In this study, 407 buildings were randomly selected as the source domain and the remaining 100 were utilized as the target domain. All the data preprocessing, modeling and experiment tasks were performed using the *R* programming language and the *Keras* package [41, 42]. In terms of missing values, variables with more than 20% missing values were removed from analysis. It is observed that the overall building power consumption distribution is right skewed, as few buildings have extremely high powers. Such measurements can be treated as data anomalies due to their rare occurrences and should be removed to avoid undesired instability in model development. A threshold-based data removal procedure was conducted in this study, i.e., the top 10% extreme cases of high powers were removed from analysis. The average power consumption patterns given different day types and primary usages are presented in Fig. 11. Evident variations in operating patterns can be observed across different building types. For instance, the daily power consumption profiles of primary and secondary schools typically rise at 6 a.m. and fall at 5 p.m., while these two timestamps are roughly 10 a.m. and 10 p.m. respectively for university classrooms. In addition, it is observed that except for university dormitories, evident pattern differences can be observed between weekdays and weekends. In general, the average power consumptions of university laboratories are the highest, while the primary and secondary school classrooms have the lowest average power consumptions.

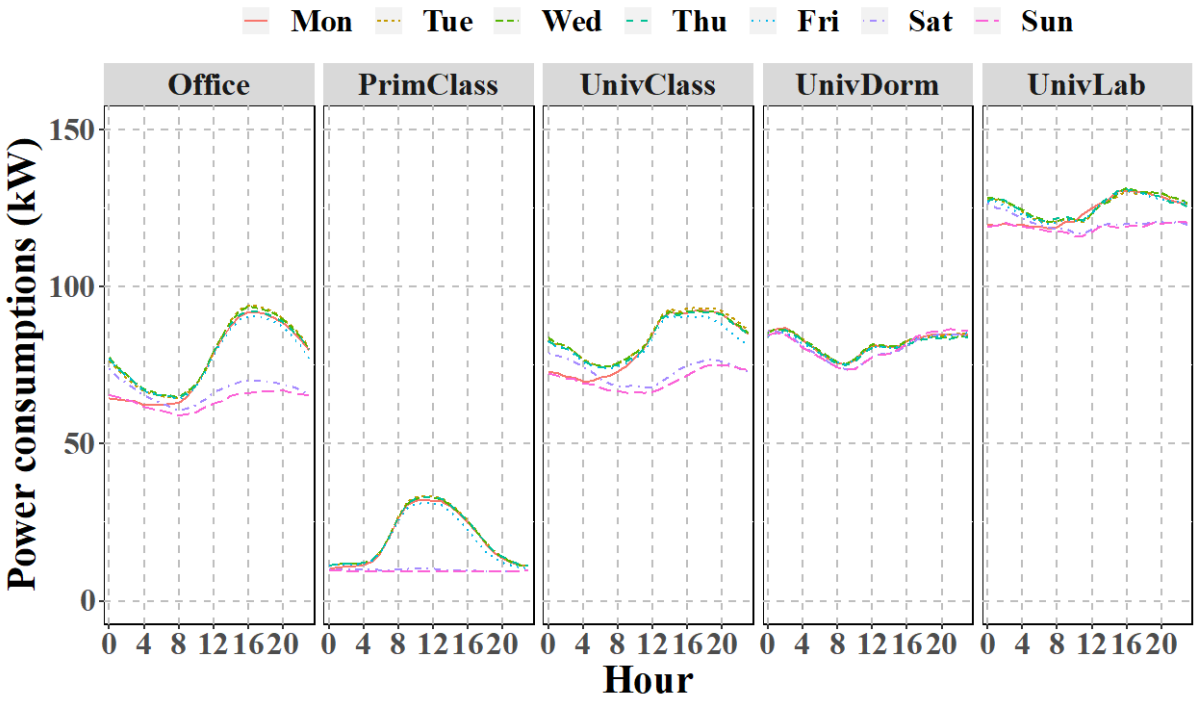

Fig.11 Average building power consumption patterns

As described in Section 3.2, the original data were transformed into suitable formats for model development. The former two inputs are three-dimensional arrays, representing historical measurements of power consumptions and outdoor conditions (i.e., dry-bulb temperature and relative humidity) in the last 7-day, and outdoor conditions for the prediction day respectively. The latter three inputs describe the *Month* (i.e., January to December), *Day Type* (i.e., Monday to Sunday) and *Primary Usage* (i.e., Office, Primary/Secondary Classrooms, University Classrooms, University Dormitories and University Laboratories). The input and output samples were generated using a daily sliding window. As shown in Table-2, 83,060 and 19,027 samples were prepared for experiments in the source and target domains respectively.

Table-2 Data samples prepared for experiments

Input and output data	Source domain	Target domain
Total sample size	83,060	19,027
Input 1: Historical measurements in last 7-day	3D array 83,060 ×168×3	3D array 19,027 ×168×3
Input 2: Outdoor conditions for the prediction day	3D array 83,060 ×24×2	3D array 19,027 ×24×2
Input 3: Month for the prediction day	2D matrix 83,060×1	2D matrix 19,027×1

Input 4: Day type for the prediction day	2D matrix 83,060×1	2D matrix 19,027×1
Input 5: Building primary usage	2D matrix 83,060×1	2D matrix 19,027×1
Output: The 24-hour ahead power consumptions	2D matrix 83,060×24	2D matrix 19,027×24

4.2 Pre-trained model optimization and performance evaluation

The data samples in the source domain are divided into training, validation and testing sets with proportions of 70%, 15% and 15% respectively. Numerical variables were standardized using the means and standard deviations calculated in the training data. Categorical variables (i.e., *Month*, *Day Type* and *Primary Usage*) were transformed into integers between 0 and *L*-1 before feeding to the embedding layer, where *L* is the number of categorical levels. For instance, Monday and Sunday are denoted as 0 and 6 respectively. As summarized in Table-1, the grid-search method was adopted to determine the optimal pre-trained model architecture. As depicted in Fig. 12, the optimal model architecture uses two 1D convolutional layers to extract local temporal features from *Input 1*. The numbers of filters are 200 and 100 respectively, each with a kernel size of 4 and a rectified linear unit (*ReLU*) activation function. The stride sizes are set as 1 and 2 respectively, representing a local feature extraction process with gradual dimensionality reductions. The convolutional operations are followed by a bidirectional recurrent layer with 48 *LSTM* recurrent units. The activation function is *Tanh*. It should be mentioned that even though *ReLU* has been reported to be successful in many deep learning tasks, it would lead to the non-convergence problem if used as the recurrent activation in this study. Both dropout and recurrent dropout techniques are 20% for model regularization. In terms of *Input 2*, one recurrent layer with 12 *LSTM* units was used. In terms of *Inputs 3*, *4* and *5*, three embedding layers are used to transform categorical values into numeric representations. The intermittent outputs from *Inputs 1* and *2* are firstly concatenated and then integrated with intermittent outputs from *Inputs 3* to *5* through element-wise multiplication. The last model component is a dense layer with 24 neurons, serving as predictions for the next 24-hour.

The generalization performance of the pre-trained model is evaluated based on the testing data in the source domain. The overall prediction performance is compared with two benchmarks: (1) using the building energy consumptions of the previous day as predictions and (2) using the building energy consumptions of the same day in previous week as predictions. Despite the simplicity of the benchmarking models, it can be quite challenging to beat due to significant

daily and weekly operating patterns. The resulting performance is reported in Table-3. The pre-trained model produces a *RMSE* of 10.97, while the first and second benchmark models lead to *RMSEs* of 16.58 and 14.39 respectively. The results validate the usefulness of the pre-trained model, as it can reduce approximately 34% (i.e., $\frac{16.58-10.97}{16.58} = 0.34$) and 24% (i.e., $\frac{14.39-10.97}{14.39} = 0.24$) of the errors produced by benchmark models. To further present the prediction performance of the pre-trained model, the *CV-RMSEs* along the 24-hour prediction horizon given different *Day Types* and *Primary Usages* are calculated and visualized in Fig. 13. In general, a slightly increasing trend in *CV-RMSEs* can be observed with the increase in prediction time steps. It is observed that *CV-RMSEs* are typically below 30% except for buildings used as primary or secondary classrooms. Such findings are identical with results reported by the authors of the Building Data Genome Project [40]. Such buildings have highly fluctuating operating schedules due to diverse occupant behaviors and therefore, the prediction accuracy can be much worse. The pre-trained model has the best performance for university laboratory buildings, as the operating schedules are relatively fixed across the whole year. The pre-trained model performs generally the same for offices, university classrooms and university dormitories in terms of *CV-RMSEs*.

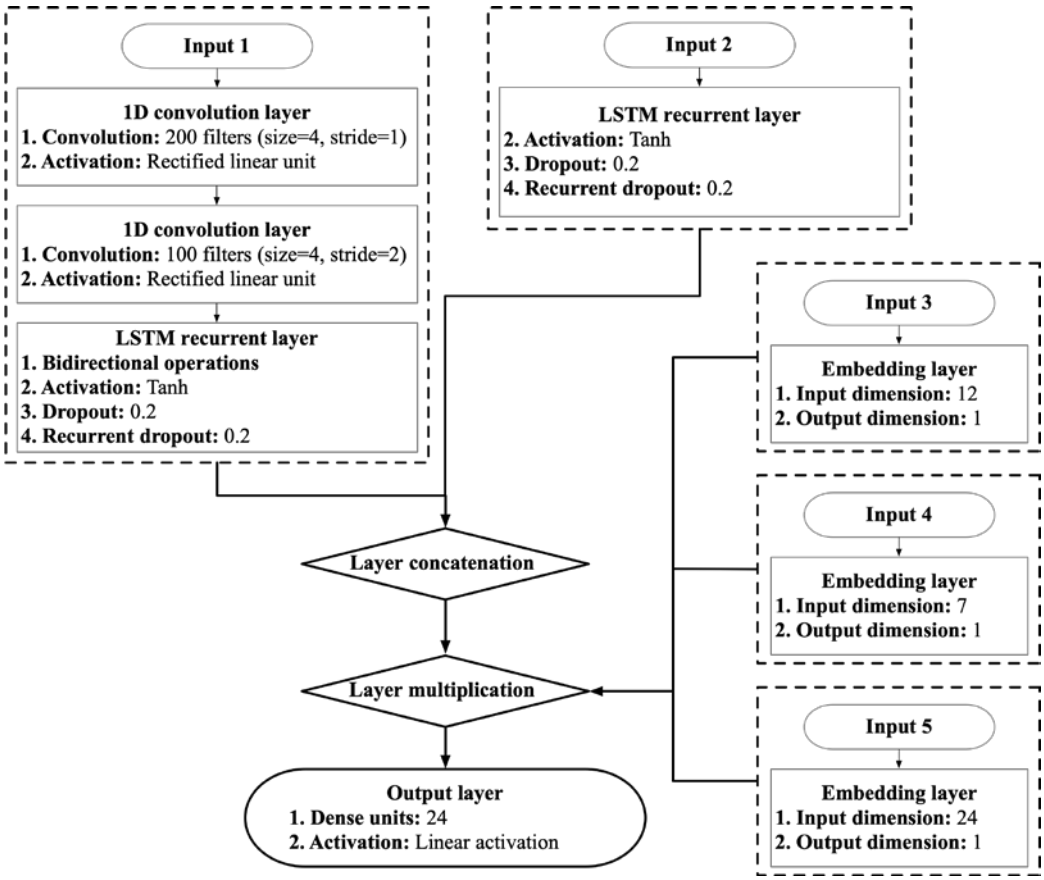

Fig. 12 The optimal pre-trained model architecture

Table-3 Prediction performance of the pre-trained model and two benchmarks

Metrics	Pre-trained model	Benchmark 1	Benchmark 2
RMSE (kW)	10.97	16.58	14.39
CV-RMSE	18.03%	27.25%	23.65%

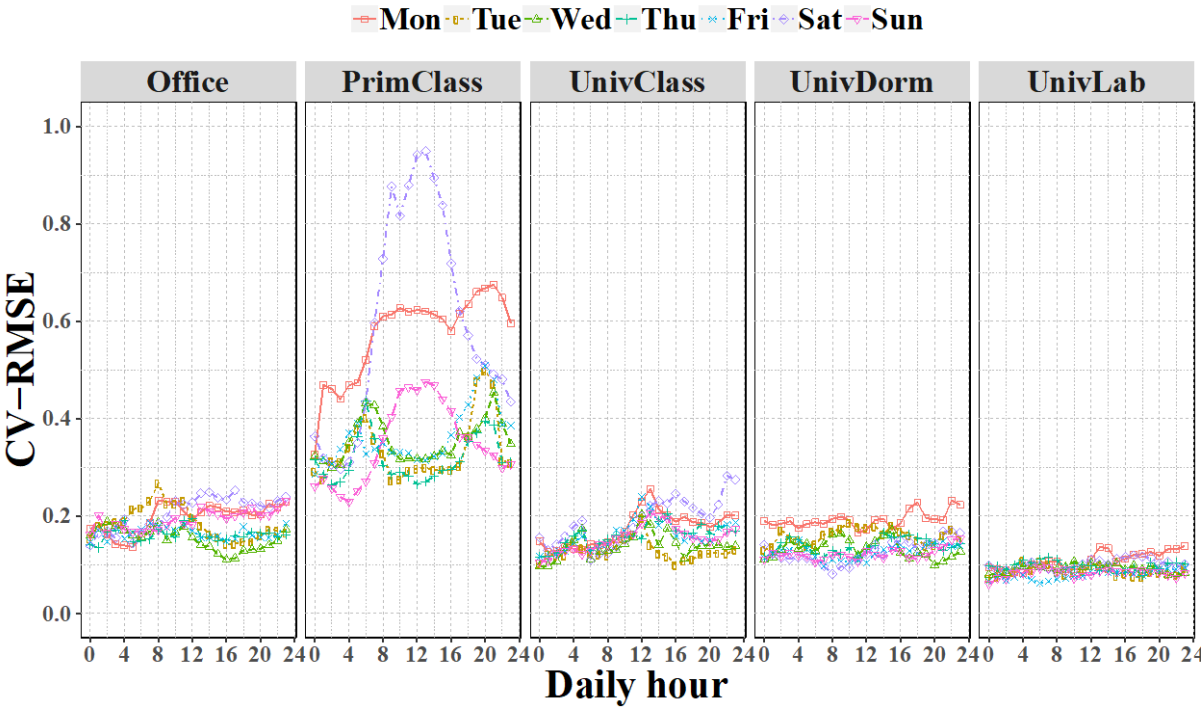

Fig. 13 The prediction performance of the pre-trained model in CV-RMSE

4.3 Evaluation of transfer learning-based building energy predictions

The values of transfer learning are evaluated based on experiments conducted for each of the 100 buildings in the target domain. As introduced in Section 3.3, two learning scenarios with different data availabilities were adopted to simulate practical applications. *Learning Scenario A* assumes data availabilities with four levels, i.e., 20%, 40%, 60% and 80%. The training data set for each building was constructed by randomly selecting samples across the whole year. Such learning scenario is in accordance with practical situations where the building has existed for a long time, yet the data available for model development is limited due to the lack of advanced building automation systems or regular data collection activities. *Learning Scenario B* assumes random month availabilities with four levels, i.e., 2-month, 5-month, 7-month and 10-month, which approximate the data proportions used in previous case. Such learning scenario is in accordance with practical situations where the building has only experienced a few operating conditions and the data collected cannot fully describe the seasonalities in building operations. The knowledge learnt by the pre-trained model is transferred to target

buildings using two implementation strategies, i.e., either for feature extraction or for weight initialization. As shown in Eq. 3, The performance of transfer learning under different data availabilities and implementation strategies is evaluated based on the performance improvement ratio (*PIR*). From a statistical perspective, the 100 buildings are randomly selected and independent to each other. The sample size is large enough to derive reliable results.

4.3.1 Experiment results in *Learning Scenario A*

Table-4 summarizes the *PIR*s calculated for 100 buildings in the target domain. As shown in Fig. 14, violin plots have been prepared as visual supplements for *PIR* distributions, where the violin widths reflect the relative frequency, the inner black points and vertical lines indicate the means and standard deviations. In general, the adoption of pre-trained models is helpful for improving the prediction performance, as *PIR*s are typical distributed above zero. It is observed that the *PIR* distribution is more uniform at 20% data availability, indicating relatively large performance variations in individual buildings. By contrast, the *PIR* distributions are more normally distributed with a mean value of around 20% at the other three data availabilities.

As shown in Table-4, a decreasing trend in *PIR* mean values can be observed with the increase in data availabilities. For instance, the *PIR* mean values are the largest at 20% data availabilities (i.e., 0.490 and 0.483 for two different implementation strategies respectively) and much lower at 60% or 80% data availabilities. The results indicate that the value of transfer learning typically decreases with the increase in training data amount. This is expected as the training data were randomly selected across the whole year and therefore, the inclusion of more training data will lead to smaller benefits of utilizing transfer learning. It should be noted that negative *PIR*s are observed across different data availabilities, indicating that the adoption of pre-trained model may lead to worse performance than developing a standalone model. Nevertheless, the proportions of negative *PIR*s are rather low. For instance, given 20% data availability, the negative *PIR* proportions are only 6.8% and 2.7% when using the pre-trained model for feature extraction and weight initialization respectively.

Even though there are differences in *PIR* mean values across different data availabilities, it is still unclear whether the difference is statistically significant or not considering the *PIR* standard deviations. A more rigorous approach is to adopt statistical hypothesis tests to address this issue. In this study, the two-sample paired t-test was adopted to quantitatively evaluate whether the *PIR* mean values are the same across different data availabilities. More specifically, the null hypothesis states that there is no difference between *PIR* mean values, while the

alternative hypothesis states the opposite. A t-statistic can be formulated as $\frac{\sum_{i=1}^{n}(PIR_{x,i}-PIR_{y,i})/n}{SD}$, where $PIR_{x,i}$ represents the *PIR* of the i^{th} building when the data availability is *x*, *n* is total building number in the target domain (i.e., 100 in this study), and *SD* is the standard deviation of *n PIR* differences. Under the null hypothesis, the t-statistic should follow a t-distribution with *n-1* degrees of freedom. Hence, considering a 99% confidence level, the difference between *PIR* mean values is statistically significant when the probability of t-test statistic is smaller than 1%. The resulting probabilities are summarized in Tables-5 and 6. To conclude, the differences between *PIR* mean values are statistically significant between 20% and the other three data availabilities, indicating that transfer learning is especially useful given limited data. The differences between *PIR* mean values are statistically insignificant in all the other pairwise comparisons, indicating that the value of transfer learning is generally the same across 40%, 60% and 80% data availabilities.

Table-4 Summary of PIR values in Learning Scenario A

Strategies	Metrics	20%	40%	60%	80%
Feature extraction	PIR means	0.490	0.222	0.150	0.162
	PIR standard deviations	0.320	0.287	0.253	0.231
Weight initialization	PIR means	0.483	0.228	0.170	0.176
	PIR standard deviations	0.294	0.260	0.191	0.182

In addition, to evaluate the superiority of the two implementation strategies in *Learning Scenario A*, the pairwise t-tests were also conducted at each data availability. None of the resulting p-values is smaller than 1%, indicating that these two implementation strategies present similar performance in terms of *PIR* mean values. Nevertheless, as shown in Table-4, the *PIR* standard deviations using the second implementation strategy are slightly smaller, indicating that using the pre-trained model for weight initialization may bring more stable performance.

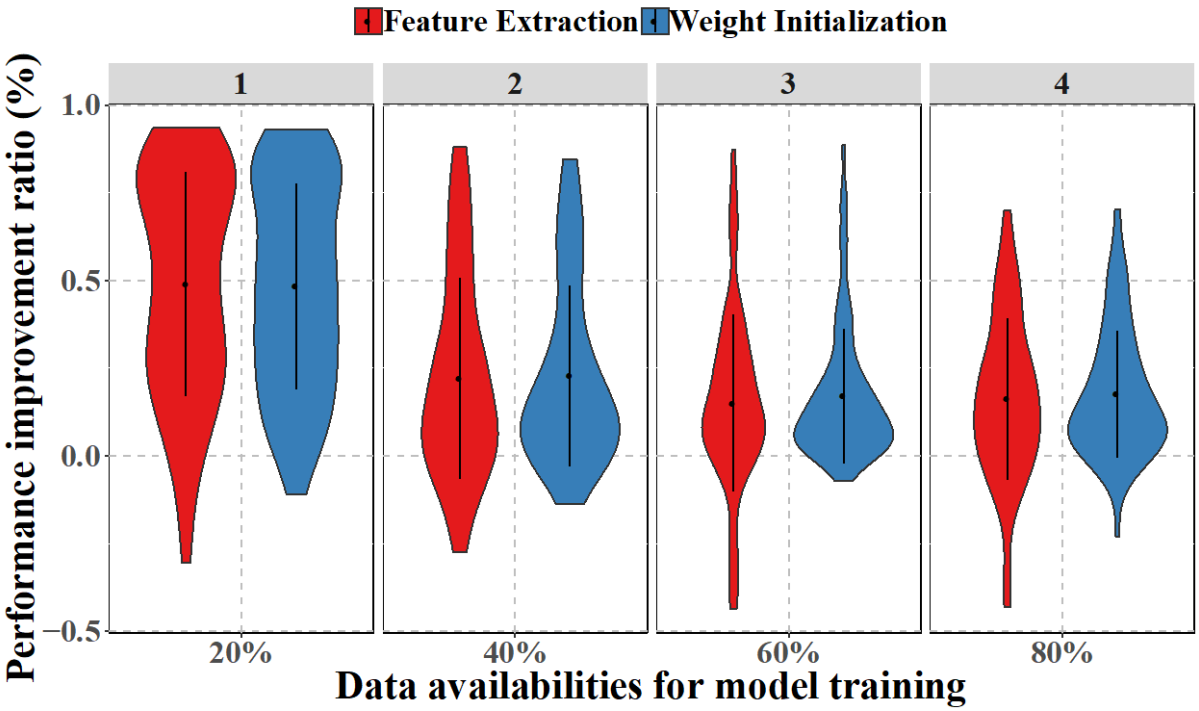

Fig.14 PIR distributions in Learning Scenario A

Table-5 *Learning Scenario A*: t-test results using implementation strategy 1

p-values	20%	40%	60%	80%
20%	NA	0.00	0.00	0.00
40%	0.00	NA	0.03	0.10
60%	0.00	0.03	NA	0.65
80%	0.00	0.10	0.65	NA

Table-6 *Learning Scenario A*: t-test results using implementation strategy 2

p-values	20%	40%	60%	80%
20%	NA	0.00	0.00	0.00
40%	0.00	NA	0.05	0.14
60%	0.00	0.05	NA	0.82
80%	0.00	0.14	0.82	NA

4.3.2 Experiment results in *Learning Scenario B*

Table-7 and Fig. 15 describe the resulting *PIR* distributions in *Learning Scenario B*. Compared with *Learning Scenario A*, the benefits of adopting the pre-trained model is much higher, i.e., the prediction error can be reduced by approximately 67.6% to 77.9% on average and there are no negative transfer cases. A slightly increasing trend in *PIR* mean values can be observed through visual inspections with the increase in data availabilities, which is contradicting to

results obtained in the *Learning Scenario A*. One possible explanation is that buildings may present completely different operating patterns in different months and therefore, the increase in training data in certain operating conditions can hardly improve the predictions in other conditions, and sometimes may lead to worse generalization performance. The results show that the knowledge learnt by the pre-trained model is particularly useful for compensating such information shortage. Therefore, transfer learning-based solutions can be very promising for new buildings which have only accumulated limited measurements from a few operating conditions.

Similar to Section 4.3.1, the two-sample paired t-test was used to quantitatively evaluate whether there are significant differences in *PIR* mean values across different data availabilities. As shown in Table-8 and Table-9, the differences in *PIR* means are statistically significant between the 2-month and the other data availabilities, while insignificant in the remaining pairwise comparisons.

In addition, the pairwise t-tests within each data availability indicates that the first implementation strategy is better in terms of *PIR* mean values, i.e., using the pre-trained model for feature extraction. Nevertheless, as shown in Table-7, the associated *PIR* standard deviations are also larger, indicating slightly more unstable performance for individual cases.

Table-7 Summary of PIR values in Learning Scenario B

Strategies	Metrics	2-month	5-month	7-month	10-month
Feature extraction	PIR means	0.729	0.774	0.779	0.777
	PIR standard deviations	0.201	0.154	0.160	0.176
Weight initialization	PIR means	0.676	0.736	0.746	0.752
	PIR standard deviations	0.168	0.121	0.130	0.125

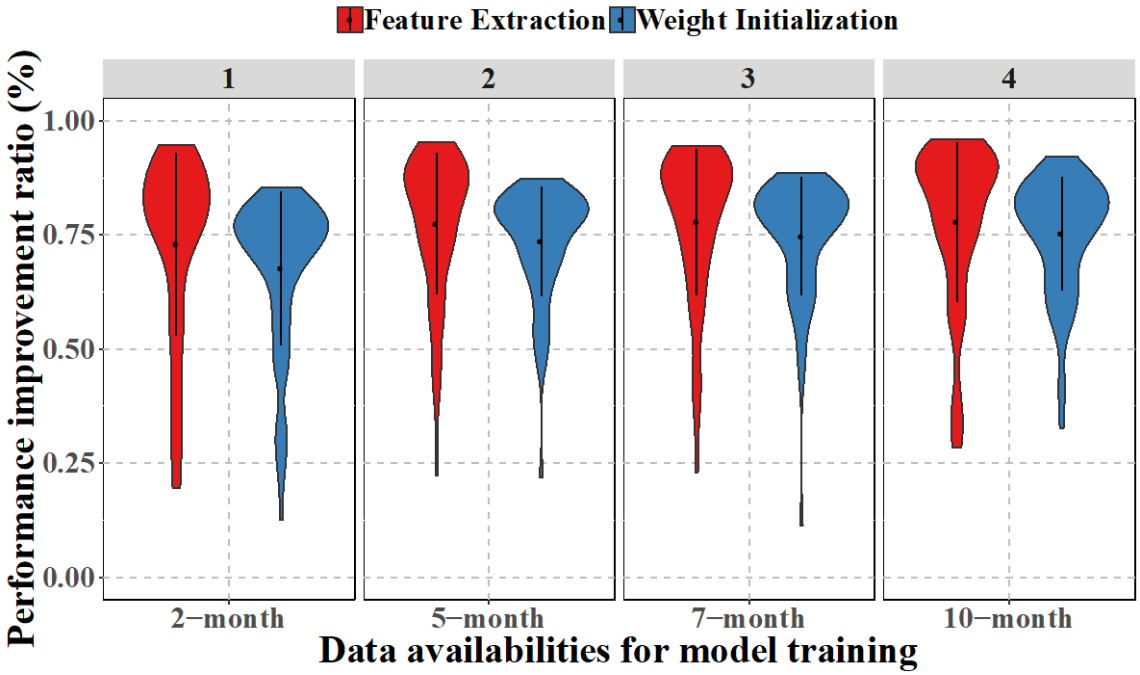

Fig. 15 PIR distributions in Learning Scenario B

Table-8 Learning Scenario B: t-test results using implementation strategy 1

p-values	2-month	5-month	7-month	10-month
2-month	NA	0.00	0.00	0.00
5-month	0.00	NA	0.17	0.17
7-month	0.00	0.17	NA	0.44
10-month	0.00	0.17	0.44	NA

Table-9 Learning Scenario B: t-test results using implementation strategy 2

p-values	2-month	5-month	7-month	10-month
2-month	NA	0.00	0.00	0.00
5-month	0.00	NA	0.04	0.05
7-month	0.00	0.04	NA	0.54
10-month	0.00	0.05	0.54	NA

5. Conclusions

The wide existence of building operational data has provided an ideal platform for developing data-driven approaches for building energy predictions. Existing solutions are mainly customized for individual buildings and highly dependent on training data amount. As a result, they may not present sufficient values for universal generalizations and practical applications. To let the building industry fully embrace the power of big data and advanced machine learning techniques, it is promising to integrate transfer learning into the knowledge discovery and application process.

This study performs a quantitative assessment of transfer learning-based methods for 24-hour ahead building energy predictions. Experiments have been designed to evaluate the usefulness of transfer learning in two learning scenarios and using different implementation strategies. The research results validate the value of transfer learning, especially when the measured data are limited and cannot fully describe seasonalities in building operating patterns. The value of transfer leaning has been quantified using the performance improvement ratio (*PIR*) and statistical tests. To summarize, in *Learning Scenario A*, the transfer learning-based methodology could reduce at least 15% of the root mean squared errors. The value of pre-trained model decreases with the increase in training data amounts. Both implementation strategies have similar performance in terms of *PIR* mean values. However, the *PIR* standard deviations indicate that more stable results can be obtained when utilizing the pre-trained model for weight initialization. In *Learning Scenario B*, the transfer learning-based methodology could reduce at least 67% of the root mean squared errors. The value of pre-trained model tends to increase with the increase in training data amounts, as the training data are drawn from certain operating conditions only. In addition, using the pre-trained model for feature extraction can lead to better performance in terms of *PIR* mean values, yet at the cost of slightly higher instability. This research validates the value of transfer learning in short-term building energy predictions. It can provide practical guidelines when developing transfer learning-based solutions for analyzing massive building operational data. Future studies will be conducted to investigate the potential of transfer learning in different building tasks (e.g., fault detection and diagnosis), using different source data availabilities (e.g., using 20% rather than 80% of the buildings as the source domain), and developing high-level strategies for knowledge transfer (e.g., taking into account the similarity between source and target buildings for knowledge transfer).

Acknowledgements

The authors gratefully acknowledge the support of this research by the National Natural Science Foundation of China (No. 51908365 and 71772125), the Philosophical and Social Science Program of Guangdong Province (GD18YGL07) and Shenzhen City (SZ2019D014), and NTUT-SZU Joint Research Program (No. 2019003).

References

[1] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 20198; 81: 1192-205.

[2] Wei YX, Zhang XX, Shi Y, Xia L, Pan S, Wu HS, Han MJ, Zhao XY. A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sustain Energy Rev 2018; 82: 1027-47.
[3] Zhao HX, Magoules F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012; 16: 3586-92.
[4] Wang ZY, Wang YR, Zeng RC, Srinivasan RS, Ahrentzen. Random forests based hourly building energy prediction. Energy Build 2018; 171: 11-25.
[5] Fan C, Sun YJ, Zhao Y, Song MJ, Wang JY. Deep learning-based feature engineering methods for improved building energy prediction. Appl Energy 2019; 240: 35-45.
[6] Fan C, Wang JY, Gang WJ, Li SH. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. Appl Energy 2019; 236: 700-710.
[7] Goodfellow I, Bengio Y, Courville A. Deep learning. 1st ed. Cambridge, London, England: MIT Press; 2016.
[8] Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data 2016; 3: 9.
[9] Pan JL, Yang Q. A survey on transfer learning. IEEE T Knowl Data En 2010; 22: 1345-59.
[10] Hermessi H, Mourali O, Zagrouba E. Deep feature learning for soft tissue sarcoma classificiation in MR images via transfer learning. Expert Syst Appl 2019; 120: 116-27.
[11] Ghiassi M, Lee S. A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. Expert Syst Appl 2018; 106: 197-216.
[12] Saeed N, King N, Said Z, Omar MA. Automatic defects detection in CFRP thermograms using convolutional neural networks and transfer learning. Infrared Phys Techn 2019; 102: 103048.
[13] Kumar S, Muhuri PK. A novel GDP prediction technique based on transfer learning using CO_2 emission dataset. Appl Energy 2019; 253: 113476.
[14] Shen S, Sadoughi M, Li M, Wang ZD, Hu C. Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries. Appl Energy 2020; 260: 114296.
[15] Chen YB, Tan HW. Short-term prediction of electric demand in building sector via hybrid support vector regression. Appl Energy 2017; 204: 1363-74.
[16] Afroz A, Urmee T, Shafiullah GM, Higgins G. Real-time prediction model for indoor temperature in a commercial building. Appl Energy 2018; 231: 29-53.
[17] Grubinger T, Chasparis GC, Natschlager T. Generalized online transfer learning for climate control in residential buildings. Energy Build 2017; 139: 63-71.

[18] Ribeiro M, Grolinger K, El Yamany HF, Higashino WA, Capretz MAM. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. Energy Build 2018; 165: 352-363.
[19] Mocanu E, Nguyen PH, Kling WL, Gibescu M. Unsupervised energy prediction in a smart grid context using reinforcement cross-building transfer learning. Energy Build 2016; 116: 646-55.
[20] Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. In: Proceedings of conferences on empirical methods in natural language processing. 2006; 120-8.
[21] Li F, Pan SJ, Jin O. Yang Q, Zhu X. Cross-domain co-extraction of sentiment and topic lexicons. In: Proceedings of the 50th annual meeting of the association for computational linguistics long papers. 2012; 1: 410-9.
[22] Tan CQ, Sun FC, Kong T, Zhang WC, Yang C, Liu CF. A survey on deep transfer learning. In: Artificial neural networks and machine learning 2018: 11141. Springer, Cham.
[23] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Advances in neural information processing systems, 2014; 27: 3320-8.
[24] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. In: Proceedings of international conference on learning representations, 2014.
[25] Li RH, Grandvalet Y, Davoine F. A baseline regularization scheme for transfer learning with convolutional neural networks. Pattern Recogn 2020; 98: 107049.
[26] Wang ZY, Srinivasan RS. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renew Sustain Energy Rev 2017; 75: 796-808.
[27] Do H, Cetin KS. Evaluation of the causes and impact of outliers on residential building energy use prediction using inverse modeling. Build Environ 2018; 138: 194-206.
[28] Guo YB, Wang JY, Chen HX, Li GN, Liu JY, Xu CL, Huang RG, Huang Y. Machine learning-based thermal response time ahead energy demand prediction for building heating systems. Appl Energy 2018; 221: 16-27.
[29] Li KJ, Xie XM, Xue WP, Dai XL, Chen X, Yang XY. A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction. Energy Build 2018; 174: 323-34.

[30] Fan C, Xiao F, Wang SW. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy 2014; 127: 1-10.
[31] Fan C, Xiao F, Yan CC, Liu CL, Li ZD, Wang JY. A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. Appl Energy 2019; 235: 1551-60.
[32] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energy 2017; 195: 222-33.
[33] Zhong H, Wang JJ, Jia HJ, Mu YF, Lv SL. Vector field-based support vector regression for building energy consumption prediction. Appl Energy 2019; 242: 403-14.
[34] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer Series in Statistics, 2nd ed. Springer; 2016.
[35] Chollet F, Allaire JJ. Deep learning with R. 1st ed. Shelter Island, New York, USA: Manning Publications; 2018.
[36] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9: 1735–80.
[37] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. 2014, arXiv: 1409.2329.
[38] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014; 15: 1929–58.
[39] Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks. In: Proceedings of NIPS 2016: 1027-35.
[40] Miller C, Meggers F. The building data genome project: An open, public data set from non-residential building electrical meters. Energy Procedia 2017; 122: 439-44.
[41] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-070; 2008. URL < http://www.R-project.org > .
[42] Allaire JJ, Chollet F and etc. Keras. Version 2.1.6. The Comprehensive R Archive Network; 2018. < https://keras.rstudio.com > .