Build Simul (2022) 15: 197–211
https://doi.org/10.1007/s12273-021-0807-6

A novel deep generative modeling-based data augmentation strategy for improving short-term building energy predictions

Cheng Fan[1,2], Meiling Chen[1,2], Rui Tang[3] (✉), Jiayuan Wang[1,2]

1. Key Laboratory for Resilient Infrastructures of Coastal Cities (Shenzhen University), Ministry of Education, China
2. Sino-Australia Joint Research Center in BIM and Smart Construction, Shenzhen University, Shenzhen, China
3. Building Technology & Urban Systems Division, Lawrence Berkeley National Laboratory, USA

Abstract

Short-term building energy predictions serve as one of the fundamental tasks in building operation management. While large numbers of studies have explored the value of various supervised machine learning techniques in energy predictions, few studies have addressed the potential data shortage problem in developing data-driven models. One promising solution is data augmentation, which aims to enrich existing building data resources for reliable predictive modeling. This study proposes a deep generative modeling-based data augmentation strategy for improving short-term building energy predictions. Two types of conditional variational autoencoders have been designed for synthetic energy data generation using fully connected and one-dimensional convolutional layers respectively. Data experiments have been designed to evaluate the value of data augmentation using actual measurements from 52 buildings. The results indicate that conditional variational autoencoders are capable of generating high-quality synthetic data samples, which in turns helps to enhance the accuracy in short-term building energy predictions. The average performance enhancement ratios in terms of CV-RMSE range between 12% and 18%. Practical guidelines have been obtained to ensure the validity and quality of synthetic building energy data. The research outcomes are valuable for enhancing the robustness and reliability of data-driven models for smart building operation management.

Keywords

building energy predictions;
data augmentation;
data-driven models;
generative modeling;
variational autoencoders

Article History

Received: 06 January 2021
Revised: 22 March 2021
Accepted: 08 April 2021

1 Introduction

The building industry is embracing the era of big data with the wide adoption of information technologies. To cope with the global trend of smart and sustainable cities, it has become increasingly promising to develop data-driven solutions for accurate and automated controls over building services systems (Wei et al. 2018). Short-term building energy predictions serve as one of the fundamental tasks in building operation management (Amasyali and El-Gohary 2018; Zhao et al. 2020). Previous studies have utilized various supervised machine learning algorithms for short-term building energy predictions (Fan et al. 2021a). A large number of single model-based methods have been developed for regression and classification problems (Yu et al. 2010; Shao et al. 2020). To further enhance the prediction performance, ensemble models, which are typically developed using bootstrap aggregating (Gong et al. 2020; Zhou et al. 2020), boosting (Walker et al. 2020; Chen et al. 2021) and stacking (Wang et al. 2020) techniques, have been proposed for building energy predictions. Compared with models of relatively shallow architectures (Seyedzadeh et al. 2019), the recent development in deep learning has encouraged researchers to utilize deep learning models for various data analytic tasks, such as unsupervised feature engineering, supervised energy predictions and semi-supervised fault detection (Fan et al. 2017; Fan et al. 2019a; Fan et al. 2021b&c). Encouraging results have been obtained in terms of prediction accuracies, data compatibilities and flexibilities (Wang and Srinivasan 2017). Nevertheless, existing studies mainly assumed that

E-mail: RuiTang@lbl.gov

List of symbols

CVAE	conditional variational autoencoder	$P(A,B)$	joint probability of A and B
CV-RMSE	coefficient of variation of root mean squared error	$P(A\|B)$	conditional probability of A given B
		PER	performance enhancement ratio
GAN	generative adversarial network	RMSE	root mean squared error
LSTM	long short-term memory	$T_1, T_2, \ldots, T_n$	time steps from 1 to n
$M_1, M_2, \ldots, M_{12}$	month from January to December	VAE	variational autoencoder

individual buildings have sufficient high-quality data, while in practice the data resource may not be satisfactory to ensure the reliability and robustness of complicated data-driven models. For instance, given the absence of advanced building automation systems or power metering systems, existing buildings may only rely on manual labors for data collection, resulting in limited data with irregular and relatively large collection intervals. Another data shortage example is for new buildings which only operate for a few weeks or months. As a result, the building operational data amounts are rather limited due to the lack of data accumulation time. In addition, the quality of building operational data may not be satisfactory considering the wide existence of malfunctions in sensor, data transmission and storage systems (Sun et al. 2020; Fan et al. 2021d). In such a case, the potential of advanced machine learning algorithms cannot be fully realized as complicated data-driven models may not generalize well due to the overfitting and non-convergence problems (Hastie et al. 2016).

There are two promising strategies to tackle the data shortage challenge, i.e., transfer learning-based and data augmentation-based strategies. The main idea of transfer learning is to utilize existing data resources from well-measured buildings to facilitate data-driven model development in poor-measured buildings (Weiss et al. 2016). Researchers have investigated the potential of transfer learning for short-term building energy predictions using neural networks, resulting in significant model improvement in various data shortage scenarios (Fan et al. 2020; Li et al. 2021). Grubinger et al. (2017) developed a transfer learning-based framework for indoor environment predictions in residential buildings. Ribeiro et al. (2018) utilized transfer learning to achieve reliable energy predictions across various building types (Ribeiro et al. 2018). Previous studies have validated the potential of transfer learning in integrating and utilizing existing building data resources. However, it is non-trivial to develop customized solutions for individual buildings considering the physical, environmental and social differences between source and target buildings.

The data augmentation-based strategy can be applied as a lightweight solution to tackle practical data shortage problems in the building field. Data augmentation refers to a set of techniques to increase the diversity of existing data by generating meaningful yet synthetic data (Goodfellow et al. 2016; Um et al. 2017). In the building field, such techniques can be applied to generate synthetic data to increase data amounts or potentially describe unseen working conditions for reliable data-driven model development. Data augmentation techniques have been successfully used in various fields for enhancing data-driven model performance, e.g., image data can be rotated or cropped for computer vision tasks while audio data can be distorted and scaled for speech recognition tasks (Chollet and Allaire 2018). Compared with transfer learning, data augmentation-based strategy is easy to implement and adaptable for individual buildings. On the one hand, it can be integrated as a data preprocessing step to enhance the quality of building data analysis results, as improvements are typically expected even given sufficient data (Goodfellow et al. 2016). On the other hand, it is extremely useful for enriching building operational data for specific building energy management tasks. For instance, the training data for system fault detection and diagnosis model development are typically imbalanced with very few faulty measurements. In such a case, data augmentation techniques can be applied to create synthetic faulty data, which helps to enhance the generalization performance of fault classification models (Rashid and Louis 2019).

Despite the presence of promising data augmentation algorithms and tools, few studies have been conducted to investigate and evaluate their potentials in building data analysis. More specifically, few researchers have addressed the necessity and value of data augmentation for time series regression tasks. Considering that building operational data are in essence time series data and typically of limited quality, it is essential to develop customized data augmentation methods to enhance the practical values of data-driven approaches in building energy management. To address this research gap, this study proposes a novel deep generative modeling-based method for short-term building energy predictions. The research outline is summarized as below.

- The main purpose of this study is to investigate the value of conditional variational autoencoders for augmenting building operational data for short-term building energy predictions.
- Data experiments, which utilizes both conventional and advanced time series augmentation techniques, have been conducted using actual energy data from 52 buildings.
- The paper has the following structure. Section 2 describes the theoretical background on data augmentation techniques. The research methodology is introduced in Section 3. Data experiment results are reported and discussed in Section 4. Conclusions are drawn in Section 5.

2 Theoretical background

2.1 Conventional data augmentation techniques for time series data

Building operational data are typically collected by automation systems with fixed time intervals, making them in essence time series data. Taking 24-hour building energy data as examples, Figure 1 illustrates classic data augmentation techniques for time series data. As shown in Figure 1(a), the first is called rotation or flipping, which simply inverts data signs while fixing magnitudes (Wen et al. 2020). Such methods can be used to simulate different sensor positions (Fawaz et al. 2018) or data orientations (Shao et al. 2019). The second is to create synthetic sequences by changing temporal data orders (Le Guennec et al. 2016). For instance, the permutation method is shown in Figure 1(b), where the original sequence is divided into three equal-length temporal segments with eight-hour time period, based on which random perturbation is performed to form a new time series. Time-warping is another popular approach for changing temporal characteristics. In such a case, synthetic data are created by stretching or shortening the original time series with different warping ratios. The third is to create synthetic data by introducing small changes in data magnitudes. Jittering and scaling are two widely used approaches. As shown in Figures 1(c) and (d), the former simulates additive sensor noises by adding Gaussian noises, while the latter changes data scales by multiplying the original data with random scalars (i.e., the scaling factor is 0.5 as shown in Figure 1(d)). It is not reasonable to introduce large variations or change temporal orders for regression problems, as they will greatly impact the temporal dependencies for time series predictions. By contrast, the jittering method can simulate malfunctions in building sensors or data collection systems. It is therefore adopted as the classic approach for augmenting building energy data in this study.

The abovementioned data augmentation techniques are relatively easy for implementation. Nevertheless, these

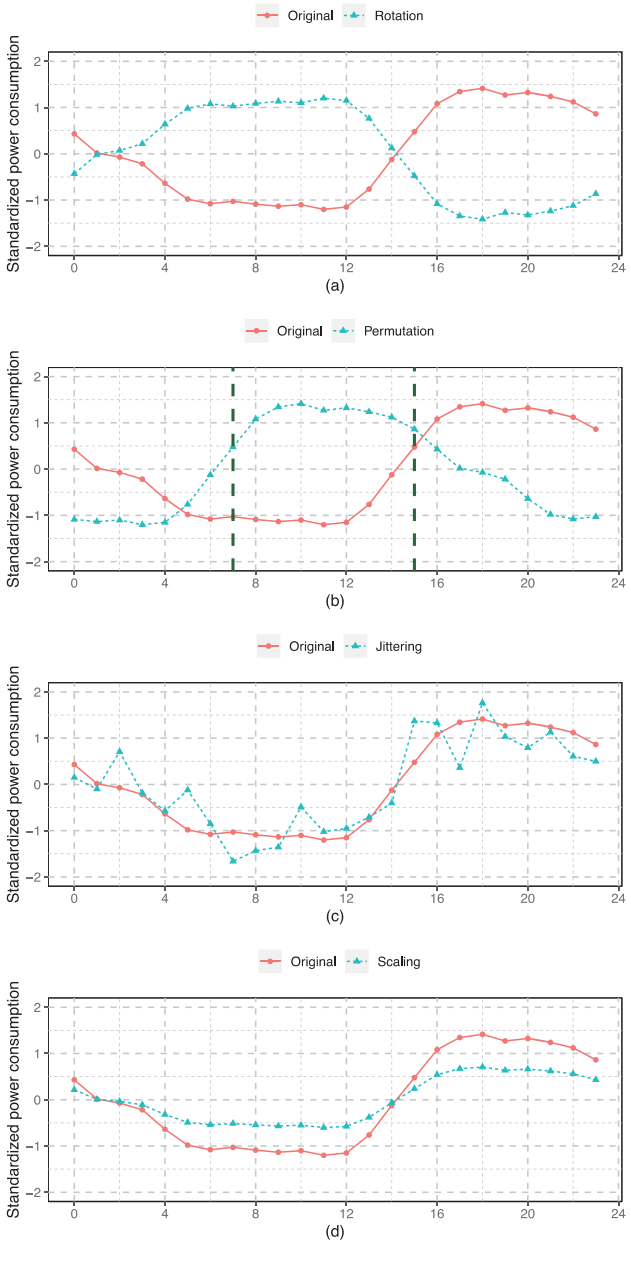

Fig. 1 Conventional data augmentation techniques for building energy data

techniques are heavily dependent on prior knowledge about the data invariance properties and the data diversity gained is rather limited (Um et al. 2017). To overcome this drawback, advanced data augmentation techniques based on generative models have been proposed. The basics are introduced in the following section.

2.2 Generative modeling-based data augmentation techniques

There are two general machine learning approaches, i.e., discriminative and generative learning (Ng and Jordan 2001).

Both have been used for regression or classification tasks. Discriminative models directly learn the conditional probability $P(Y|X)$ from data, where X and Y denote model input and output variables respectively. By contrasts, the generative models learn the joint probability $P(Y|X)$, based on which the Bayes rule can be applied for generating predictions. Existing studies have shown that discriminative models are more efficient and effective in predictive modeling, while the joint probability learned by generative models can be valuable for other purposes, e.g., generating new samples for data augmentation (Antoniou et al. 2018). Compared with typical data augmentation techniques, generative models can produce synthetic data with broader variations and higher quality and thus, providing more useful information to enhance the generalization performance of discriminative models (Goodfellow et al. 2016). It should be mentioned that such performance enhancement cannot be simply achieved by using more powerful or advanced discriminative models due to the difference in their learning paradigms.

The rapid development in deep learning has provided powerful tools for generative modeling. Two deep learning-based generative models, i.e., generative adversarial networks (GANs) and variational autoencoders (VAEs), have gained great popularity due to their excellence in capturing complicated data distributions (Frid-Adar et al. 2018; Simão et al. 2019; Xu et al. 2019). The main intuition of generative modeling-based data augmentation techniques is to map the real data into a set of latent distributions, based on which synthetic data can be sampled during data generation. A GAN model, which consists of a generator and discriminator, is trained in an adversarial way to generate high-quality synthetic data. Previous studies have shown GAN models were capable of producing highly realistic daily patterns describing both general trends and stochastic dynamics in building operations (Wang and Hong 2020). However, GAN models can be extremely difficult to train in practice and typically require extensive experiments to find optimal model parameters (Creswell et al. 2017; Tian et al. 2019). By contrasts, VAEs are much easier to train using gradient-based methods and have obtained state-of-the-art results in generative modeling (Goodfellow et al. 2016; Bregere and Bessa 2020). More importantly, VAEs are capable of learning structured latent spaces, which can provide more controls over the synthetic data generation process (Chollet and Allaire 2018). Therefore, this study adopts VAEs as the main technique for augmenting building energy data. The technical details are shown in Section 2.3.

2.3 Basics on variational autoencoders

Variational autoencoders (VAEs) are developed in 2013 (Kingma and Welling 2013). VAEs are variations of conventional autoencoders, which aim to reconstruct original data through an encoding and decoding process. As illustrated in Figure 2, a conventional autoencoder has a bottleneck architecture and consists of an encoder and a decoder. The encoder transforms the original data into a lower dimensional latent vector Z. The decoder tries to reconstruct the original data by taking Z as inputs. Various training constraints can be applied to derive meaningful latent representations of original data. Autoencoders have been widely used for data dimensionality reduction and feature engineering (Baldi 2012).

As shown in Figure 3, instead of simply compressing the original data into a latent vector, the encoder of VAE transforms the original data into a set of means and variances (i.e., denoted as μ and σ^2 respectively), which specify the characteristics of normal distributions (Um et al. 2017). Random sampling is then performed to obtain a latent vector Z from latent normal distributions, where $Z = \mu + \sigma \odot \epsilon$, ϵ are random values drawn from standard normal distributions, and $\odot$ denotes element-wise product. Such random sampling process, which is known as the reparameterization trick, enables the gradient backpropagation through the whole VAE models while ensuring the stochasticity (Goodfellow et al. 2016). VAEs are trained with the aim of minimizing two types of losses. The first is reconstruction loss, which compares the difference between original and synthetic data. The second is regularization loss, which aims to reduce the overfitting risk while developing a well-structured latent space (Chollet and Allaire 2018). It should be mentioned that the dimension of latent vectors should be optimized to ensure the quality of synthetic data.

In practice, it is often desired to generate synthetic data given certain conditions. As an example, buildings may have different operation patterns in different months or day types (e.g., weekdays and weekends). Rather than random

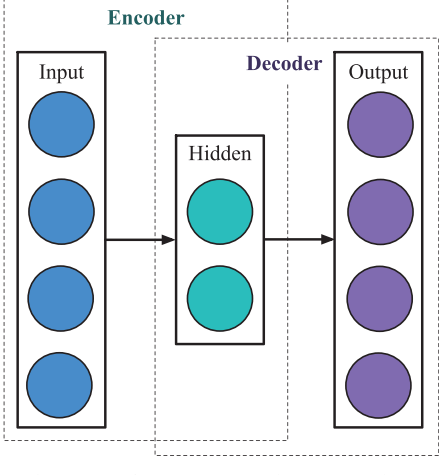

Fig. 2 The schematic of a conventional autoencoder

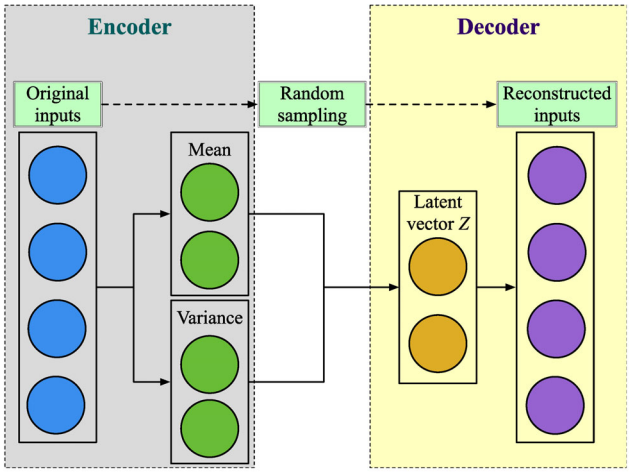

Fig. 3 The schematic of a variational autoencoder (VAE)

data augmentation, one may want to generate synthetic data for a given month or day type. Conditional variational autoencoders (CVAE) can be applied to address such needs (Sohn et al. 2015). As shown in Figure 4, conditional information, which can be represented as either continuous or one-hot encoding vectors, is fed to both encoder and decoder as inputs. Once converged, the CVAE decoder can be applied to generate synthetic data given certain conditions. To summarize, CVAE models can provide users with more controls over the synthetic data generation process and therefore are used for augmenting building energy data in this study.

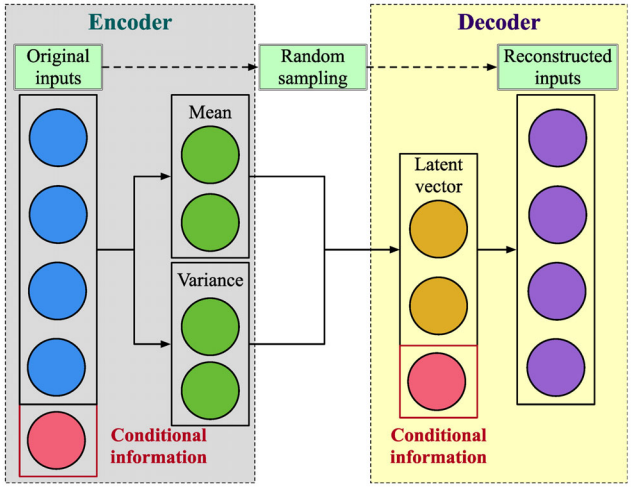

Fig. 4 The schematic of a conditional variational autoencoder (CVAE)

3 Research methodology

3.1 Research outline

As shown in Figure 5, this research aims to investigate and evaluate the potential of data augmentation for building energy analysis. To ensure the research validity, a set of buildings with different types and energy scales have been adopted for data experiments. The prediction task is defined as a multi-step building energy prediction with a prediction horizon of 24-hour. Data experiment are conducted separately for each building to evaluate the usefulness of data augmentation strategies. The training, validation and testing data are randomly selected with proportions of 70%, 15% and 15%, respectively. The training and validation data are used for model training and validation, while the generalization performance is evaluated based on testing data. Two data augmentation strategies are utilized to generate synthetic data based on training data. The first is to create synthetic data through the classic jittering method, i.e., adding random Gaussian noises. The second generates synthetic data based on CVAEs, which are developed through fully connected and one-dimensional (1D) convolutional neural networks. Prediction models are then developed for each building under different data augmentation settings. The value of data augmentation strategies is evaluated based on accuracy metrics on testing data.

3.2 Data augmentation strategies for building energy data

As an initial step, building operational data are prepared into suitable formats for 24-hour ahead building energy predictions. Considering the practical data availability, the prediction model consists of two types of input variables. The first is the *Month* of the prediction day, which can be used to describe seasonalities in building operations and indoor occupancy schedules. The second are historical building energy data. Building operations have significant weekly and daily patterns (Fan et al. 2019b; Piscitelli et al. 2021). Therefore, the maximal time lag (i.e., denoted as w) for model inputs is defined as one week. For instance, given hourly data collection interval, the building energy data at time step T–167, T–166, …, T will be used to predict energy consumptions at T+1, T+2, …, T+24. Once the data format is determined, data partitioning is performed to randomly divide the data for each building into three segments, i.e., training, validation and testing data, each with probabilities of 70%, 15% and 15%, respectively. The training data are used for model development. The validation data are used to calculate the validation loss when applying the early-stopping training scheme. The model performance is evaluated using the testing data. It should be mentioned that the main research scope is to explore the value of data augmentation strategies for short-term building energy predictions. Future studies can be conducted considering different training data availabilities.

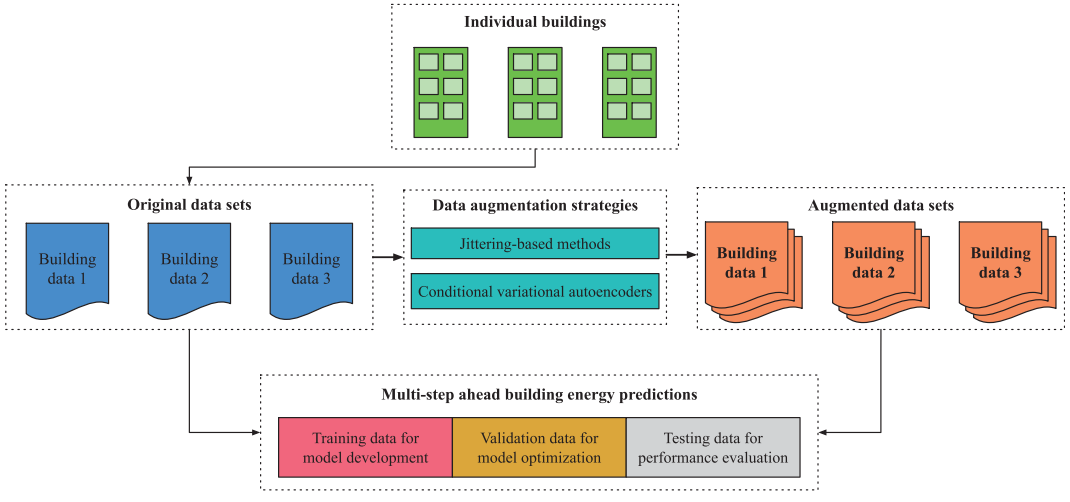

Fig. 5 Research outline

Different data augmentation strategies are then applied to enrich the training data. The first is the classic jittering method, which injects random noises drawn from Gaussian distributions to the original energy data. The magnitudes of random noises can be controlled by setting different standard deviations for Gaussian normal distributions. In this study, different levels of standard deviations and the amount of synthetic data are specified for data experiments. The second strategy utilizes CVAEs for synthetic data generation. As shown in Figure 6, two types of CVAE models (i.e., denoted as CVAE-1 and CVAE-2) are developed based on fully connected and 1D convolutional layers respectively. While CVAE-1 treats each value in the 8-day building energy pattern as independent, CVAE-2 utilizes 1D convolutional operations to capture temporal dependencies among successive energy measurements. Considering that buildings typically have different operation patterns given different seasons and occupancy schedules, the time variable *Month* is transformed into one-hot encoders as conditional information. It should be mentioned that other variables can be integrated as the conditional information for more customized synthetic data generations. For instance, building operations typically present dramatic differences between weekdays and weekends and hence, the *Day Type* can be used as additional conditional information. In this study, the prediction model takes 7-day historical data as inputs, which in essence have already include the information of *Day Type* for the prediction day. Therefore, to make the CVAE and prediction model more concise, only *Month* is adopted as the conditional information for synthetic data generation.

CVAE models are trained based on training data only. The early-stopping training scheme is adopted to prevent the overfitting problem and the patience was set as 20. More specifically, the model loss on validation data will be

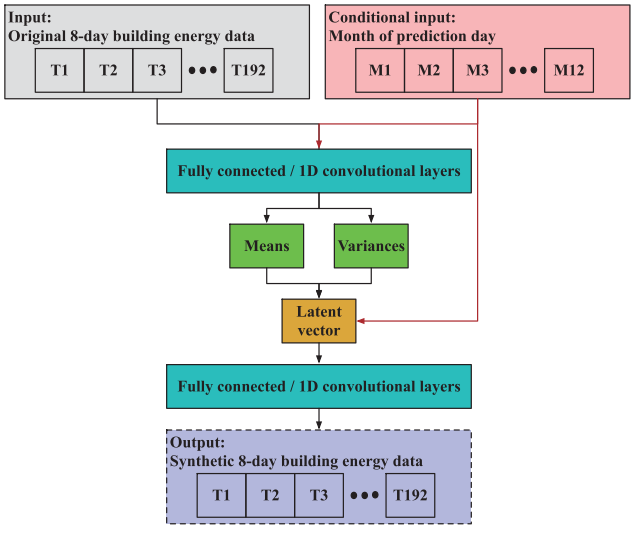

Fig. 6 CVAEs for synthetic building energy data generation

calculated at each training iteration. If the validation loss does not decrease for over 20 iterations, the training process will terminate. As described in Section 2.3, the latent vector dimensions can significantly affect the quality of synthetic data. Therefore, besides general model architectures, the latent size should also be optimized for individual buildings.

3.3 Predictive modeling on 24-hour ahead building energy consumptions

There are three main approaches for multi-step ahead building energy predictions, i.e., the recursive, direct, and multi-input multi-output approaches (Fan et al. 2019c). Our previous work has shown that the direct approach, which develops separate models for each time steps, can achieve better performance as it can effectively avoid the error accumulation problem observed in the recursive approach

while resulting in more accurate predictions compared with the MIMO approach (Gal and Ghahramani 2016). Therefore, this study adopts the direct approach for developing short-term building energy prediction models.

The artificial neural network is selected as the modeling technique. The main considerations are two-fold. Firstly, artificial neural networks provide great flexibilities in model architectures. It can be easily adaptable for implementing the direct approach by changing the number of output neurons. In such a case, there is no need to develop completely different models for each time step, resulting in potential reductions in computational costs. Secondly, through the use of various convolutional and recurrent operations, artificial neural networks can accurately capture temporal data dependencies, resulting in excellent performance in time series predictions (Hochreiter and Schmidhuber 1997; Gal and Ghahramani 2016). Therefore, this study selects neural networks for short-term building energy predictions. Prediction models are developed based on one-dimensional convolutional and recurrent operations. Optimizations in terms of the model architecture will be performed to ensure the generalization performance.

3.4 Performance evaluation

To ensure the result validity, this study adopts buildings with different types and scales for analysis. For each building, prediction models are developed using the training data only, with or without data augmentation. The early-stopping scheme is adopted to prevent the overfitting problem considering the mean square error in validation data. The generalization performance is evaluated based on the remaining testing data. The root mean squared error (RMSE) and the coefficient of variation of root mean squared error (CV-RMSE) are selected for performance comparisons as shown in Eqs. (1) and (2), where the total sample size is denoted as n, $\hat{y}$ and y denote predicted and actual values, respectively. The performance enhancement ratios (PERs) are defined to quantify the usefulness of data augmentation strategies in short-term building energy predictions. As shown in Eq. (3), PER is defined as the relative reduction in RMSEs when data augmentation is used, where $RMSE_1$ and $RMSE_2$ are obtained by prediction models based on the original and augmented data sets, respectively. On the premise that data augmentation is useful, $RMSE_2$ should be smaller than $RMSE_1$, resulting in positive PER. On the contrary, PER should be negative if data augmentation has negative influence on short-term building energy prediction.

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

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

$$\text{PER} = \frac{\text{RMSE}_1 - \text{RMSE}_2}{\text{RMSE}_1} \tag{3}$$

4 Data experiment results and discussions

4.1 Data description

The Building Data Genome Project was utilized for data experiment (Miller and Meggers 2017). The project provides one-year building operational data for 507 non-residential buildings. All the data analysis tasks were performed using the R programming language (R Development Core Team 2008). Three general types of information are available for analysis. The first describes the general building information, including building physical attributes (e.g., locations, total floor areas and occupancy numbers) and main functionalities (i.e., office, primary/secondary classrooms, university classrooms, university dormitories and university laboratories). The second provides ten-minute interval descriptions on outdoor conditions using a set of meteorological variables, such as outdoor temperature and relative humidity. The third is one-year building energy consumption data collected at hourly interval.

As mentioned in Section 3.2, the maximal time lag for model inputs was set as 168 considering weekly seasonalities in building operations. Therefore, the building energy data were firstly merged with outdoor conditions and then transformed into subsequences with a length of 192 (i.e., 168 + 24 = 192) for each building. The data subsequences were generated with a daily stride, i.e., 24-hour. To ensure the fairness and validity of data experiments, each testing building should possess the same amount of data subsequences for model development and evaluation. However, the buildings in the Building Data Genome Project have different data availabilities due to the presence of missing values or outliers. It is observed that 52 buildings (i.e., 10 offices, 12 university classrooms, 7 university dormitories and 23 university laboratories) have exactly the same amount of data subsequences (i.e., 276). Given larger data subsequence numbers, the number of testing buildings selected would be much smaller than 30, which may not justify the statistical significance of experiment results. Therefore, these 52 buildings have been adopted for data analysis. The general energy patterns of these 52 buildings are depicted in Figure 7 according to their building types using median values. Significant differences between weekdays and weekends

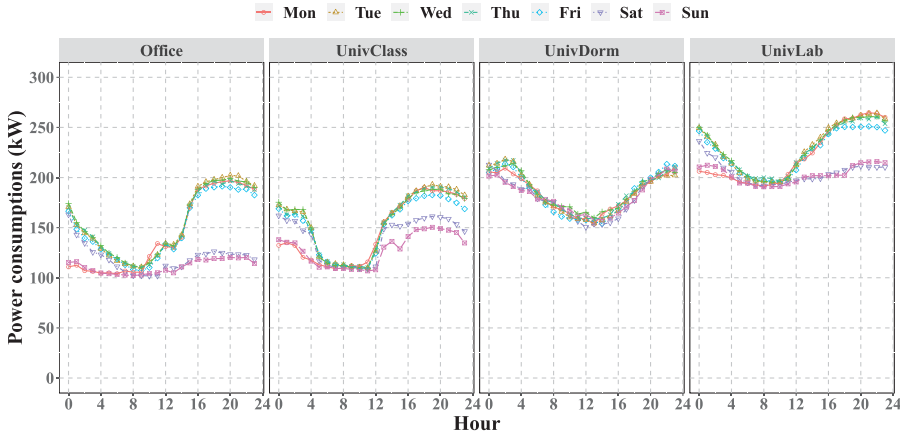

Fig. 7 Typical building energy patterns for different building types

can be observed for all building types except for university dormitories. This is expected as the occupancy schedule of university dormitories is relatively fixed across the whole week. In addition, university laboratories have slightly higher energy uses, which is also expected due to the operation of energy-intensive equipment.

4.2 Synthetic data generation

As described in Section 3.2, data experiments have been conducted for each building separately. The available data subsequences of individual buildings were randomly partitioned into training, validation and testing data sets with proportions of 70%, 15% and 15%, respectively. The data augmentation was performed based on the training data set only. To evaluate the impact of synthetic data amounts to short-term building energy predictions, two sets of synthetic data have been generated and denoted as *5-fold* and *10-fold* respectively. The *5-fold* data have 965 subsequences, which is 5 times the number of training subsequences (i.e., $276 \times 0.70 = 193$) for individual buildings. The *10-fold* data have 1930 subsequences, which is 10 times the number of training subsequences for individual buildings.

Two data augmentation strategies have been implemented for each building. As shown in Figure 8, the first applies random Gaussian noises to each time step of the training data subsequences. Three levels of standard deviations were set to control the magnitudes of Gaussian noises, i.e., 1%, 5% and 10% of the standard deviation in original building power consumption data.

The second develops CVAEs for synthetic data generation. CVAE-1 was developed using fully connected layers. As shown in Table 1, the grid-search has been used for model parameter optimization. To summarize, CVAE-1 was designed with symmetric architecture, indicating that both hidden layer and hidden neuron sizes would be the same for encoder and decoder models. To reduce computation costs in model optimization, only two essential model parameters were optimized, i.e., the numbers of hidden layers (i.e., 0, 1 or 2) and neuron sizes (i.e., 50, 100, 150).

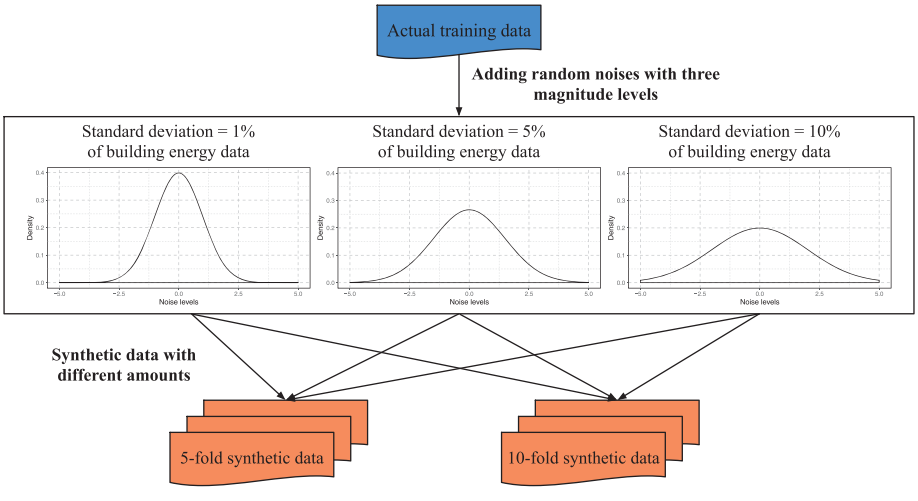

Fig. 8 Synthetic data generation using jittering-based method

Table 1 Grid-search settings for CVAE development

Type	Model parameters	Candidates
CVAE-1	The number of fully connected layers	0, 1, 2
	Hidden neuron numbers at each layer	50, 100, 150
CVAE-2	The number of 1D convolutional layers	1, 2, 3, 4, 5
	The filter number at each layer	50, 100, 150

The conditional information, i.e., *Month* of the prediction day, was transformed into one-hot representations and integrated with other inputs using the concatenation approach. Rectified linear units (*ReLU*) were used as the activation function except for the encoder and decoder output layers, where *Linear* activation function was used. The candidates for latent size were 2, 5 and 10, and the optimal value may vary for different buildings. To capture temporal dependencies in time series data, CVAE-2 was developed using 1D convolutional layers and optimized using a similar grid-search fashion. To reduce computation costs in model optimization, only the numbers of hidden layers and filters were optimized, as these two typically have the largest impacts on model performance. There were five candidate hidden layer values for the encoder and decoder, i.e., one to five. The filter number was optimized considering three candidate values (i.e., 50, 100 and 150) and was set equal for all 1D convolutional layers. Once optimized, synthetic data could be created for each building using CVAE decoders. For each building, synthetic data were generated considering two synthetic data amounts, i.e., *5-fold* and *10-fold*. As shown in Figure 9, the conditional information, i.e., *Month* of the prediction day, was generated using random uniform distribution and the latent vectors were randomly sampled from standard normal distributions. As an example, Figure 10 depicts the boxplot between

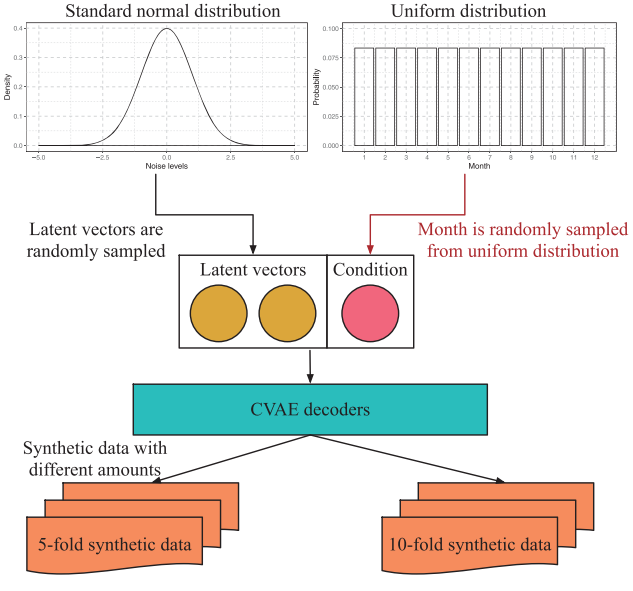

Fig. 9 Synthetic data generation using CVAE-based methods

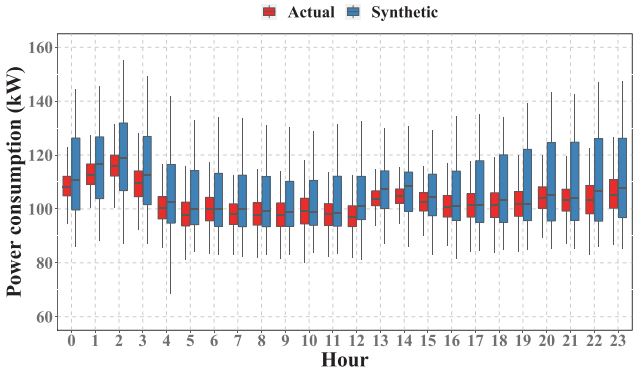

Fig. 10 An example boxplot of synthetic and actual energy data for a university dormitory building

synthetic and actual building energy data for a university dormitory building. It is observed that the mean values of different *Hour* are approximately the same, while the synthetic data present much wider variations. In such a case, the synthetic energy data can be used to describe unseen working conditions and thereby, leading to possible enhancements in data-driven prediction models.

4.3 Prediction model development and optimization

As described into Section 3.3, the direct approach based on artificial neural networks has been adopted for 24-hour ahead building energy predictions. To ensure the fairness in evaluating data augmentation strategies, the model architecture has been optimized based on building operational data from 10 random selected buildings and fixed for all the other data experiments. The grid-search settings and results for model optimization are summarized in Table 2. Considering that building energy data present significant temporal and seasonal patterns, the model was designed with possible 1D convolutional and recurrent operations. More specifically, the model was optimized considering four essential parameters, i.e., the number of 1D convolutional layers, the filter number in 1D convolutional layers, whether to use bidirectional operations for recurrent operations, and the recurrent unit number in the LSTM layer. The recurrent unit number was designed with four candidate values (i.e., 24, 48, 72 and 96) to capture possible interactions among daily energy patterns.

The optimal model architecture is shown in Figure 11.

Table 2 Grid-search settings for prediction model optimization

Model parameters	Grid-search candidates	Results
1D convolutional layer numbers	1, 2, 3	2
The filter number	50, 100, 150, 200	100
Bidirectional operations	Yes/No	Yes
LSTM neuron numbers	24, 48, 72, 96	48

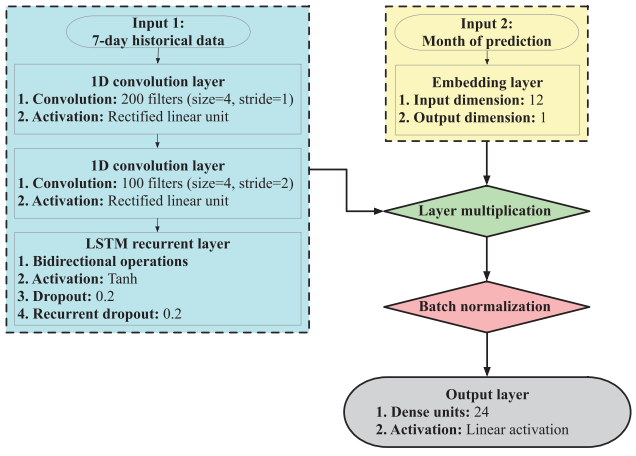

Fig. 11 The optimized prediction model architecture

The model inputs consist of one-week historical energy consumptions and the *Month* of the prediction day. Both convolutional and recurrent operations have been utilized to handle the historical energy consumptions. Two one-dimensional convolutional layers have been used to extract high-level temporal features and reduce computation costs. The filter sizes were set as 200 and 100 respectively. The *ReLU* is selected as the activation function to reduce the risk of exploding or vanishing gradients. The relationships among temporal features were then captured based on bidirectional and long short-term memory (LSTM) operations. The bidirectional operations were adopted to enrich the temporal information for recurrent model development. More specifically, a bidirectional recurrent layer consists of two recurrent operations, which take a forward and reversed pass through the input sequence respectively. The hyperbolic tangent is selected as the LSTM activation function, as other choices may lead to failures in model training. Besides the early-stopping training scheme, the dropout and recurrent dropout for recurrent operations were both set as 20% to reduce the risk of overfitting. Meanwhile, the *Month* is transformed using embedding operations and integrated with recurrent outputs using dot multiplication. Afterwards, batch normalization layer is applied to stabilize the training process. The output layer is a 24-neuron dense layer with the *Linear* activation function, representing predictions for the next 24-hour.

4.4 Evaluations and discussions on data augmentation strategies

4.4.1 Evaluations on jittering-based data augmentation strategies

Different prediction models have been developed for each building to evaluate data augmentation methods. A baseline model was firstly developed using the real data alone. The early stopping training scheme was adopted to prevent the risk of overfitting. The CV-RMSEs on testing data sets of all 52 buildings were calculated to indicate the baseline prediction performance. Figure 12 serves as the density plot of 52 CV-RMSEs. The mean and median CV-RMSEs are 13.86% and 12.04%, respectively. It is observed that the majority of CV-RMSEs are below 30%, which are in accordance with results obtained in other studies using the Building Genome Project (Miller and Meggers 2017). As explained in Section 3.4, the CV-RMSEs of baseline models were then used to calculate performance enhancement ratios (PERs) to quantify the value of different data augmentation methods.

Prediction models were then trained based on augmented data, where actual training data were combined with synthetic data for model development. As the initial attempt, six augmented data sets are generated using the jittering-based methods. Each data set varies in terms of the synthetic data amount (i.e., denoted as *5-fold* and *10-fold*) and the level of random Gaussian noises (i.e., denoted as 1%, 5% and 10%). The PERs can be calculated for each augmented data and the resulting PER distributions are summarized in Table 3. More specifically, four statistical features are reported, i.e., the mean, the 5%, 50% and 95% quantiles of PERs obtained for 52 buildings. The mean and median values are close to zero, indicating that the jittering-based method does not provide reliable enhancements for building energy predictions. Based on the 5% and 95% quantiles, it is observed that the increase in synthetic data amounts can lead to slightly better PERs, while the variations in noise levels do not present evident changing trends for PERs. As shown in Figure 13, the violin and boxplots have been adopted for visualizing the PER distributions. The violin plot illustrates the PER density while the boxplot presents statistical characteristics of PER distributions, such as the 5%, 25%, 50%, 75% and 95% quantiles. It shows that the overall performance is not satisfactory as negative PERs can be observed in almost half of the 52 buildings. It is therefore

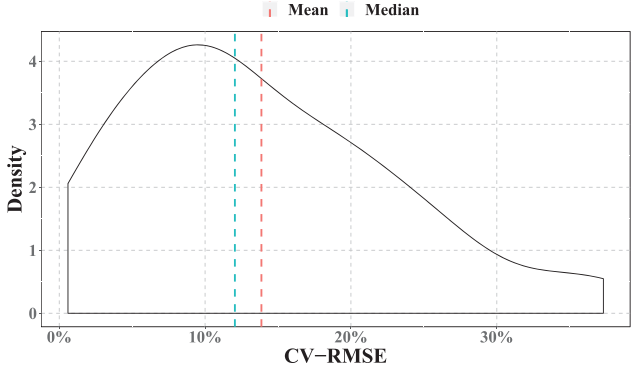

Fig. 12 The overall baseline model performance in terms of CV-RMSE

Table 3 Summary on PERs using the jittering-based strategy

Noise levels	Data folds	5% quantile	Median	Mean	95% quantile
1%	5	−0.46	0.00	−0.06	0.29
	10	−0.43	0.00	−0.03	0.35
5%	5	−0.37	0.00	−0.01	0.32
	10	−0.31	0.01	0.02	0.36
10%	5	−0.38	0.01	−0.04	0.25
	10	−0.39	0.01	−0.01	0.34

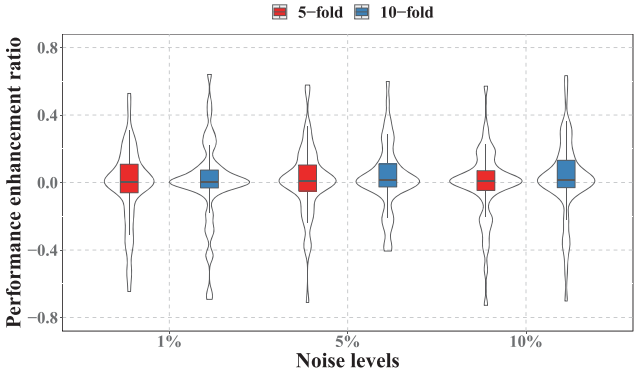

Fig. 13 PER distributions using the jittering-based data augmentation strategy

not suggested to apply such techniques for augmenting building energy data for short-term building energy predictions in practice. The data experiment results show that even though such technique has been reported useful for classification problems in other fields, it is not suitable for regression problems in the building field.

4.4.2 Evaluations on CVAE-based data augmentation strategies

The CVAE-based data augmentation methods have been applied to enrich the training data for building energy predictions. In total, four augmented data sets have been created using two types of CVAE models (i.e., denoted as CVAE-1 and CVAE-2, which utilize fully connected and 1D convolutional layers respectively) and synthetic data amounts (i.e., *5-fold* and *10-fold*).

As shown in Figure 14, the majority of PERs are positive, indicating that the augmented data are helpful in enhancing 24-hour ahead building energy predictions. Similar to Table 3, four statistical features of PERs are shown in Table 4 to summarize the performance of CVAEs for short-term building energy predictions. The results show that CVAE-1 has slightly better performance than CVAE-2 in terms of higher PER mean values and smaller standard deviations. It indicates that fully connected layers have sufficient capability in extracting high-level features in building energy data. The differences in synthetic data folds do not

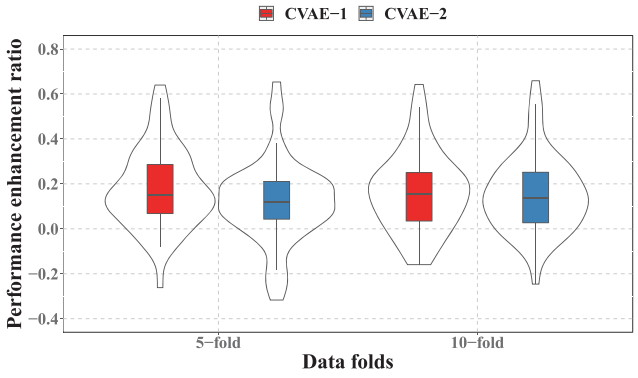

Fig. 14 PER distributions using different CVAEs and synthetic data amounts

Table 4 Summary on PERs using the CVAE-based strategy

Type	Data folds	5% quantile	Median	Mean	95% quantile
CVAE-1	5	−0.06	0.15	0.18	0.51
	10	−0.11	0.16	0.17	0.49
CVAE-2	5	−0.21	0.12	0.12	0.44
	10	−0.07	0.14	0.15	0.49

present evident impacts on PERs. One possible explanation is that the intrinsic variations of individual building operation patterns are rather limited compared with data measurements in other fields. Hence, the increase in synthetic data amount may not bring extra benefits for building energy predictions. It should be mentioned that this study assumes that 70% of the one-year hourly measurements are available for model development, which may be sufficient for reliable model development. Further studies can be conducted to explore the value of data augmentation with different training data availabilities.

Considering that there are intrinsic differences in energy patterns of different building types, the values of data augmentation in short-term building energy predictions may also vary. As shown in Figure 15, the boxplots of PERs present wider distributions for university dormitories and laboratories, while the distributions are much narrower for offices and university classrooms. It indicates that the potential value of data augmentation in energy predictions can be higher for buildings with relatively fixed energy patterns (e.g., university dormitories and laboratories). Besides, similar to previous findings, there is no significant PER differences when synthetic data are created using different CVAE models or of different amounts.

Figure 16 presents the relative frequency of optimized CVAE latent dimensions in different experiment settings. In general, a positive correlation can be observed between latent dimensions and their relative frequencies in optimized models. Most of the CVAE models select ten as the

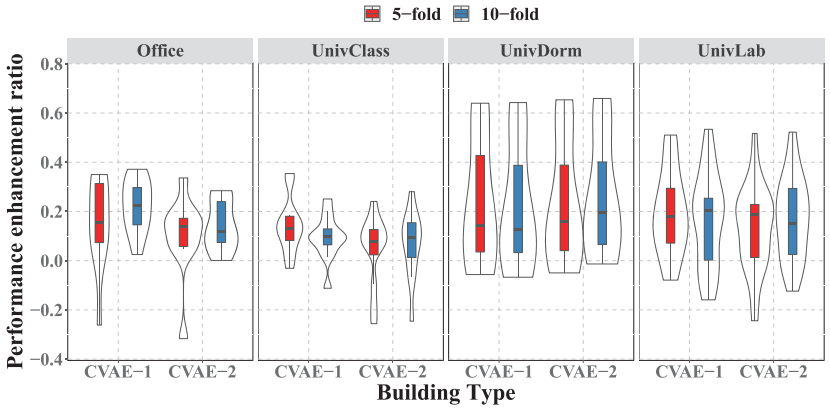

Fig. 15 PER distributions given different building types

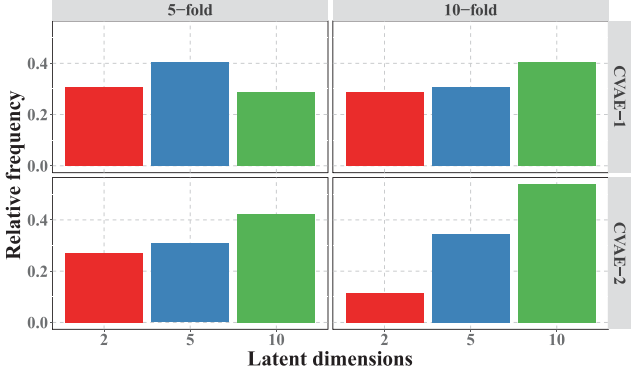

Fig. 16 The relative frequency of optimized CVAE latent dimensions

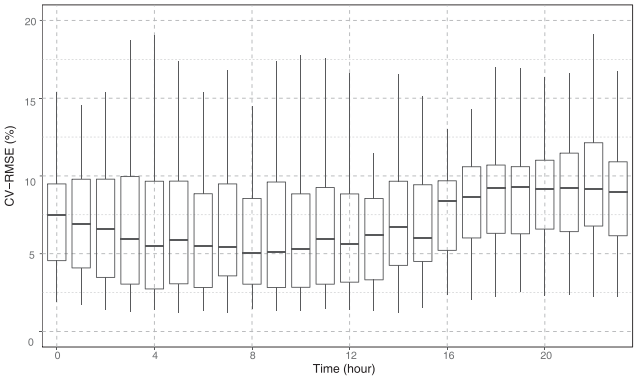

Fig. 17 The baseline model performance along the 24h prediction horizon

optimized latent dimension, except for the first experiment settings where five has the highest relative frequency. The results indicate that the quality of synthetic data may become better with the increase in latent dimensions, which in turns leads to better performance in short-term building energy predictions. This is expected as a larger latent dimension typically leads to more information for synthetic data generation.

4.4.3 General performance along the 24-hour prediction horizon

In-depths investigation have been performed to assess the prediction performance along the 24-hour prediction horizon. Figure 17 presents CV-RMSE boxplots at each time step using the baseline model. In general, CV-RMSEs during office hours are smaller than those during non-office hours. This is expected as the occupancy schedules of office hours are typically more fixed and present less random variations, making it much easier for energy predictions.

Figure 18 illustrates the average CV-RMSEs at each time using prediction models developed based on different data augmentation strategies. The findings are in accordance with those reported in previous two sections. The red solid line represents the average baseline CV-RMSE at each time step. It is surround by CV-RMSE lines of different jittering methods, indicating that the jittering-based data augmentation strategy does not provide consistent improvement in building energy predictions. By contrasts, the CVAE-based data augmentation methods lead to relatively considerable CV-RMSE reductions at each time step, especially during non-office hours. One possible explanation is that CVAE models can generate synthetic samples with meaningful random variations, which helps to cover the unseen input space in training data. Experiment results also indicate that fully connected CVAEs (i.e., CVAE-1) can lead to slightly better prediction performance than 1D convolutional CVAEs (i.e., CVAE-2). It indicates that 1D convolutional operations alone may not effectively extract temporal dependencies in building energy data. As a possible direction for performance improvement, further studies may consider the integrated use of 1D convolutional and fully connected layers or recurrent operations for developing CVAE models.

5 Conclusions

To fully realize the potential of advanced machine learning

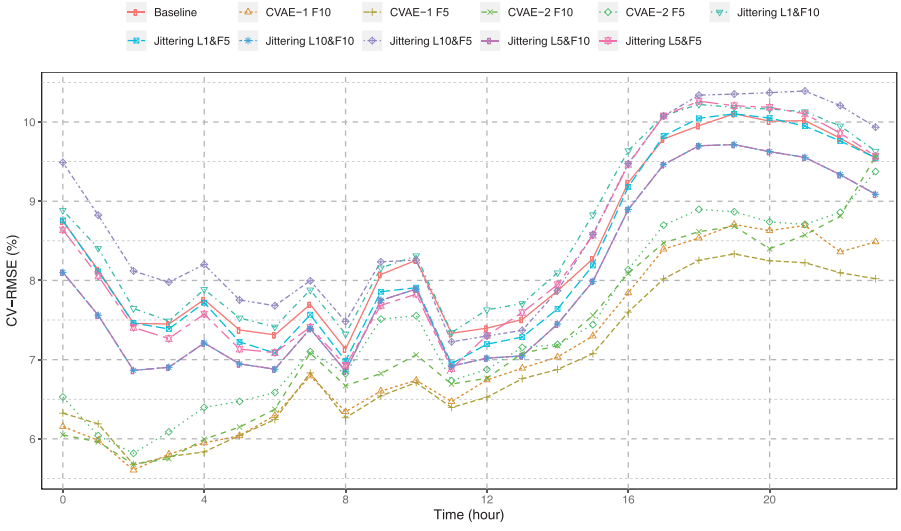

Fig. 18 The average CV-RMSEs using different data augmentation methods

techniques, it is essential to provide sufficient high-quality training data to avoid the underfitting or over-fitting problem during model development. To tackle practical data shortage problems and enhance data-driven model reliabilities, this study proposes a novel generative modeling-based data augmentation method for building energy data. Considering the time series nature of building energy data, two types of conditional variational autoencoder (CVAE) have been designed to generate synthetic yet potentially meaningful data for model development. The usefulness of data augmentation has been tested in the task of 24-hour ahead building energy predictions. Data experiments have been conducted using 52 buildings to validate and quantify the value of methodology proposed. A novel metric, i.e., performance enhancement ratio or PER, has been defined to quantify the value of synthetic data in building energy predictions. The results indicate that CVAE-based generative learning methods can effectively enhance the performance of short-term building energy predictions. The major result findings are as below:

- The average PERs ranges from 12% to 18% when CVAE-based methods are used for augmenting building energy data.
- CVAE models with fully connected layers are sufficient to generate useful synthetic building energy data for reliable model development.
- The potential value of data augmentation in energy predictions can be higher for buildings with relatively fixed energy patterns (e.g., university dormitories and laboratories).
- The classic time series augmentation method, i.e., jittering, do not have consistent performance for building energy predictions, as the average and median PERs are close to zero.

The study has provided practical guidelines and insights for augmenting building energy data, based on which advanced data analytics can be developed to facilitate data-driven tasks in smart building energy managements. The method is of particular use when the data collection interval is relatively large. In such a case, synthetic data can be used to enrich data representativeness, which typically lead to improvements in model generalization performance. Future studies can be conducted from two perspectives. The first is to address the problem of synthetic data quality evaluation. Rather than using indirect approaches, direct approaches can be developed to ensure the flexibility and applicability of data augmentation, e.g., comparing distributions between synthetic and actual data or constructing classification models for data authenticity checks. The second is to investigate the power of data augmentation for other data-driven tasks in building operation management, e.g., tackling the imbalanced data problem in fault detection and diagnosis tasks.

Acknowledgements

The authors gratefully acknowledge the support of this research by the National Natural Science Foundation of China (No. 51908365, No. 71772125) and the Philosophical and Social Science Program of Guangdong Province, China (GD18YGL07).

References

Amasyali K, El-Gohary NM (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81: 1192–1205.

Antoniou A, Storkey A, Edwards H (2018). Data augmentation generative adversarial networks. arXiv: 1711.04340v3.

Baldi P (2012). Autoencoders, unsupervised learning and deep architectures. *JMLR Workshop and Conference Proceedings*, 27: 37–50.

Bregere M, Bessa RJ (2020). Simulating tariff impact in electrical energy consumption profiles with conditional variational autoencoders. *IEEE Access*, 8: 131949.

Chen Z, Xu P, Feng F, et al. (2021). Data mining algorithm and framework for identifying HVAC control strategies in large commercial buildings. *Building Simulation*, 14: 63–74.

Chollet F, Allaire JJ (2018). Deep Learning with R. New York: Manning Publications.

Creswell A, White T, Dumoulin V, et al. (2017). Generative adversarial networks: An overview. In: Proceedings of IEEE Signal Processing Magazine Special Issue on Deep Learning for Visual Understanding.

Fan C, Xiao F, Zhao Y (2017). A short-term building cooling load prediction method using deep learning algorithms. *Applied Energy*, 195: 222–233.

Fan C, Sun Y, Zhao Y, et al. (2019a). Deep learning-based feature engineering methods for improved building energy prediction. *Applied Energy*, 240: 35–45.

Fan C, Xiao F, Yan C, et al. (2019b). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Applied Energy*, 235: 1551–1560.

Fan C, Wang J, Gang W, et al. (2019c). Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. *Applied Energy*, 236: 700–710.

Fan C, Sun Y, Xiao F, et al. (2020). Statistical investigations of transfer learning-based methodology for short-term building energy predictions. *Applied Energy*, 262: 114499.

Fan C, Yan D, Xiao F, et al. (2021a). Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches. *Building Simulation*, 14: 3–24.

Fan C, Liu X, Xue P, et al. (2021b). Statistical characterization of semi-supervised neural networks for fault detection and diagnosis of air handling units. *Energy and Buildings*, 234: 110733.

Fan C, Liu Y, Liu X, et al, (2021c). A study on semi-supervised learning in enhancing performance of AHU unseen fault detection with limited labeled data. *Sustainable Cities and Society*, 70: 102874.

Fan C, Chen M, Wang X, et al. (2021d). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, 9: 652801.

Fawaz HI, Forestier G, Weber J, et al. (2018). Data augmentation using synthetic data for time series classification with deep residual networks. arXiv: 10808.02455v1.

Frid-Adar M, Klang E, Amitai M, et al. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. In: Proceedings of IEEE 15th International Symposium on Biomedical Imaging.

Gal Y, Ghahramani Z (2016). A theoretically grounded application of dropout in recurrent neural networks. In: Proceedings of NIPS.

Gong M, Wang J, Bai Y, Li B, Zhang L (2020). Heat load prediction of residential buildings based on discrete wavelet transform and tree-based ensemble learning. *Journal of Building Engineering*, 32: 101455.

Goodfellow I, Bengio Y, Courville A (2016). Deep Learning. Cambridge, MA, USA: MIT Press, USA.

Grubinger T, Chasparis GC, Natschläger T (2017). Generalized online transfer learning for climate control in residential buildings. *Energy and Buildings*, 139: 63–71.

Hastie T, Tibshirani R, Friedman J (2016). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. New York: Springer.

Hochreiter S, Schmidhuber J (1997). Long short-term memory. *Neural Computation*, 9: 1735–1780.

Kingma DP, Welling M (2013). Auto-encoding variational Bayes. arXiv: 1312.6114.

Le Guennec A, Malinowski S, Tavenard R (2016). Data augmentation for time series classification using convolutional neural networks. In: Proceedings of ECML/PKDD Workshop in Advanced Analytics and Learning on Temporal Data.

Li A, Xiao F, Fan C, et al. (2021). Development of an ANN-based building energy model for information-poor buildings using transfer learning. *Building Simulation*, 14: 89–101.

Miller C, Meggers F (2017). The Building Data Genome Project: An open, public data set from non-residential building electrical meters. *Energy Procedia*, 122: 439–444.

Ng AY, Jordan MI (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In: In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01).

Piscitelli MS, Brandi S, Capozzoli A, et al. (2021). A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings. *Building Simulation*, 14: 131–147.

R Development Core Team (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rashid KM, Louis J (2019). Times-series data augmentation and deep learning for construction equipment activity recognition. *Advanced Engineering Informatics*, 42: 100944.

Ribeiro M, Grolinger K, El Yamany HF, et al. (2018). Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. *Energy and Buildings*, 165: 352–363.

Seyedzadeh S, Rahimian FP, Rastogi P, et al. (2019). Tuning machine learning models for prediction of building energy loads. *Sustainable Cities and Society*, 47: 101484.

Shao S, Wang P, Yan R (2019). Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry*, 106: 85–93.

Shao M, Wang X, Bu Z, et al. (2020). Prediction of energy consumption in hotel buildings via support vector machines. *Sustainable Cities and Society*, 57: 102128.

Simão M, Neto P, Gibaru O (2019). Improving novelty detection with generative adversarial networks on hand gesture data. *Neurocomputing*, 358: 437–445.

Sohn K, Yan X, Lee H (2015). Learning structured output representation using deep conditional generative models. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15).

Sun Y, Haghighat F, Fung BCM (2020). A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings*, 221: 110022.

Tian C, Li C, Zhang G, et al. (2019). Data driven parallel prediction of building energy consumption using generative adversarial nets. *Energy and Buildings*, 186: 230–243.

Um TT, Pfister FMJ, Pichler D, et al. (2017). Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In: Proceedings of ACM International Conference on Multimodal Interaction.

Walker S, Khan W, Katic K, et al. (2020). Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. *Energy and Buildings*, 209: 109705.

Wang R, Lu S, Feng W (2020). A novel improved model for building energy consumption prediction based on model integration. *Applied Energy*, 262: 114561.

Wang Z, Hong T (2020). Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN). *Energy and Buildings*, 224: 110299.

Wang Z, Srinivasan RS (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, 75: 796–808.

Wei Y, Zhang X, Shi Y, et al. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, 82: 1027–1047.

Weiss K, Khoshgoftaar TM, Wang D (2016). A survey of transfer learning. *Journal of Big Data*, 3: 9.

Wen Q, Sun L, Song X, et al. (2020). Time series data augmentation for deep learning: A survey. arXiv: 2002.12478v1.

Xu P, Du R, Zhang Z (2019). Predicting pipeline leakage in petrochemical system through GAN and LSTM. *Knowledge-Based Systems*, 175: 50–61.

Yu Z, Haghighat F, Fung BCM, et al. (2010). A decision tree method for building energy demand modeling. *Energy and Buildings*, 42: 1637–1646.

Zhao Y, Zhang C, Zhang Y, et al. (2020). A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment*, 1: 149–164.

Zhou Y, Chen J, Yu ZJ, et al. (2020). A novel model based on multi-grained cascade forests with wavelet denoising for indoor occupancy estimation. *Building and Environment*, 167: 106461.