REVIEW

National Science Open
3: 20230068, 2024
https://doi.org/10.1360/nso/20230068

Engineering

Special Topic: Energy Systems of Low Carbon Buildings

Novel machine learning paradigms-enabled methods for smart building operations in data-challenging contexts: Progress and perspectives

Cheng Fan[1,2,3], Yutian Lei[2,3], Jinhan Mo[1,2,3], Huilong Wang[1,2,3,*], Qiuting Wu[2,3] & Jiena Cai[2,3]

[1]*Key Laboratory for Resilient Infrastructures of Coastal Cities, Ministry of Education, Shenzhen University, Shenzhen 518067, China;*
[2]*Sino-Australia Joint Research Center in BIM and Smart Construction, Shenzhen University, Shenzhen 518067, China;*
[3]*College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518067, China*

*Corresponding author (email: wanghuilong@szu.edu.cn)

Received 30 October 2023; Revised 12 January 2024; Accepted 31 January 2024; Published online 2 February 2024

Abstract: The increasing availability of building operational data has greatly encouraged the development of advanced data-driven technologies for smart building operations. Building operational data typically suffer from data quality problems, such as insufficient labeled and imbalanced data, making them incompatible with conventional machine learning algorithms. Recent advances in data science have provided novel machine learning paradigms to tackle such data challenges for practical applications, such as transfer learning, semi-supervised learning, and generative learning. This review aims to present the progress and perspectives on the effective utilization of novel machine learning paradigms for three major building energy management tasks, i.e., building energy predictions, fault detection and diagnosis, and control optimizations. In-depth discussions have been provided to demonstrate the pros and cons of different learning approaches in terms of data compatibility, modeling difficulties, and possible application scenarios, which could be helpful for enhancing the feasibility of data-driven technologies for smart building operations.

Keywords: smart building operations, building energy management, transfer learning, semi-supervised learning, generative learning

INTRODUCTION

The energy-efficient building operations are of utmost significance to achieve energy saving and sustainable development. Conventional building operation technologies are heavily dependent on domain expertise and human labors, making them ineffective and unscalable for large-scale and automated implementations. The rapid development of internet of things (IoT) and data infrastructures has enabled the collection of massive amounts of building operational data, which provides solid foundations to develop data-driven technologies for practical applications. Leveraging advances in data science and machine learning, data-driven technologies have the potential to effectively explore hidden knowledge in massive building operational data to facilitate daily operation decision-makings. Assuming the availability of sufficient and high-quality data, existing studies have validated the use of data-driven technologies in various energy management tasks, such

© The Author(s) 2024. Published by Science Press and EDP Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

as building thermal load predictions [1], fault detection and diagnosis [2], indoor environment monitoring and inferences [3], thermal comfort evaluations [4], demand-response and optimal controls over air conditioning systems [5]. Compared with traditional human-centric approaches, data-driven approaches are highly promising to enhance the accuracy and automation levels in building system operations.

Despite the promising potential, data-driven applications in the building field are still somehow limited due to the violation of perfect data assumption in practice. Admittedly, the wide adoption of building IoTs is generating large amounts of real-time measurements with shorter collection intervals, e.g., a few seconds to minutes. However, such data are typically of intrinsic limitations, making them incompatible with conventional data learning paradigms. Taking the data representativeness as an example, building operations typically present strong seasonalities, which call for relatively long data accumulation time to ensure data representativeness. For new buildings, it may not be possible to collect sufficient data in short time to cover all possible operating conditions. As a result, the data-driven model developed may not generalize well due to less representative training data. As we shall demonstrate in later sessions, similar data challenges exist due to the high costs of human labor, the privacy issues in cross-building data sharing, the intrinsic characteristics of building system operations, etc. Advanced data analytics are urgently needed to fully uncover the potential of building operational data for reliable and intelligent building operations.

The recent development in mathematics, statistics, and computer science has devised novel machine learning paradigms to tackle various data quality challenges, such as transfer learning [6], semi-supervised learning [7], generative learning [8], reinforcement learning [9], etc. The main intuition is to effectively utilize existing resources to derive robust data-driven solutions, which in turn may greatly improve the feasibility of machine learning algorithms in data-scarce contexts. Such machine learning paradigms have been successfully applied in various industries, handling complex and multi-source information ranging from structured time series data [10] to unstructured image or textual data [11]. A natural question arises as whether such novel machine learning paradigms are useful in the building field. In this regard, this paper aims to present the progress and perspectives on the effective utilization of novel machine learning paradigms on three major building energy management tasks, i.e., building energy predictions, fault detection and diagnosis, and control optimizations. In this review, representative studies, which were mainly published in the past five years with keywords related to building energy management applications and machine learning paradigms, have been selected and reviewed to present different types of ideas and approaches used to tackle the building data challenges. It should be noted that reinforcement learning has gained increasing popularity in building optimal controls, and several review papers [12,13] do exist to provide the general picture, and therefore, are not included in this study. In-depth discussions and possible research directions are illustrated, which may be helpful for upgrading data analytics for reliable and robust building operations.

TYPICAL BUILDING OPERATION TASKS AND DATA CHALLENGES

Three typical building operation tasks

In general, there are three typical building operation tasks, i.e., building energy predictions, fault detection and diagnosis, and control optimization [14]. The first task, building energy predictions include predictive modeling tasks on building power or energy consumptions [15], cooling and heating loads [16]. Taking the

building cooling load prediction as an example, conventional supervised learning algorithms can be readily applied by selecting influencing variables as inputs, including time variables (e.g., Month and Hour) serving as proxy variables describing seasonalities and occupant schedules, environmental variables (e.g., temperature and relative humidity) characterizing operating conditions, and previous cooling load measurements depicting historical operating behaviors [17]. Such predictions can be conducted at either building-, system- or component-levels and are mainly used as foundations to detect anomalies in building operations [18,19] or devise optimal control strategies to enhance operating efficiencies [20,21].

The second task is fault detection and diagnosis, aiming to identify the operating status of building systems or components while locating possible causes and severity levels of different faults [22,23]. Conventional methods mainly use domain expertise to develop rule-based methods for fault inferences [24]. Alternatively, data-driven methods can be developed using machine learning algorithms to distinguish between various operating statuses. It can be formulated as either an unsupervised or a supervised task. For instance, assuming the majority of the operating data are normal operations, unsupervised clustering analysis algorithms, such as *k*-means or density-based spatial clustering of applications with noise (DBSCAN), can be applied to find clusters corresponding to normal or faulty operations [25,26]. By contrast, regression models can be developed to predict the values of monitoring variables under normal conditions, based on which the differences or residuals between predicted and actual measurements are used for fault inferences [27,28]. As a more straightforward solution, classification models can be developed to classify the label or operating status of each data sample to be either normal or faulty [29,30].

The third task is control optimization, which intends to determine the optimal settings to ensure energy efficiency at given operating conditions [31]. In practice, control optimization consists of three major steps. The basis is to formulate an objective function consisting of both controllable and uncontrollable variables under certain constraints. Afterwards, prediction models, developed using either physical or data-driven methods, are developed to describe the energy use and functionalities of different components. Finally, mathematical methods such as linear programming [32], or heuristic methods such as genetic algorithms and particle swarm optimization [33,34], can be applied to solve the optimization task. Control optimizations can be conducted either at component or system levels to achieve local or global optimizations. Taking chillers as an example, local optimization can be performed by maximizing the energy efficiency or coefficient of performance (i.e., COP) at given cooling demands. In such a case, COP prediction models can be developed using the part-load ratio, the supplied and returned temperatures of chilled and condensing water, and the associated water flowrates as inputs [35]. In theory, local optimization methods will try to increase the supplied chilled water temperature and decrease the returned condensing water temperature to achieve higher chiller COPs. The resulting control strategy may not be optimal for the water-side system, as it will increase the power consumption in chilled water pumps and cooling towers [36].

Major data challenges in building operations

In general, there are three major data challenges for efficiently and effectively analyzing the time-series data collected from building operations, i.e., the lack of data representativeness, insufficient labeled data, and imbalanced data.

As we have described in previous sections, the lack of data representativeness may occur when the data

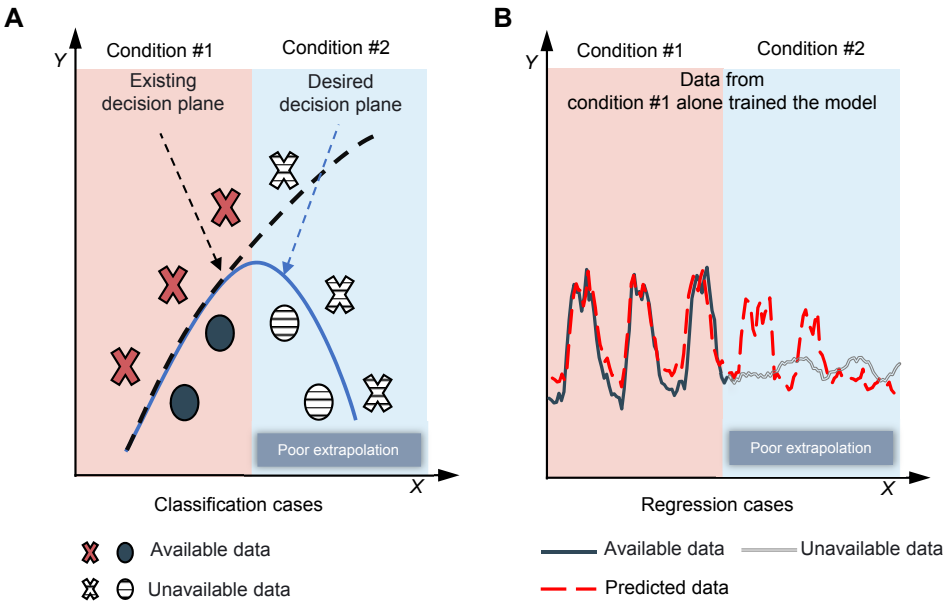

Figure 1 An illustration of poor data-driven model performance due to limited data representativeness.

collected cannot fully describe seasonal or periodic building operating conditions. As shown in Figure 1, it is particularly true for new buildings with short data collection periods or buildings with unitary operating modes. One main limitation of nearly all data-driven models is the relatively poor extrapolation ability, i.e., data-driven models will become less valid when making inferences outside the data region with sufficient training data [37]. In essence, the lack of data representativeness may force data-driven models to make extrapolating predictions for unseen building operating conditions and thereby, leading to poor generalization performance. As an example, assuming that building energy prediction models are developed using operating data collected in hot and humid outdoor environments, they are unlikely to generate reliable predictions for the unseen cold and dry outdoor conditions.

As shown in Figure 2, the availability of insufficient labeled data is another major challenge for developing data-driven models. In the building field, data labels may refer to the ground truth or actual operating conditions of each measurement. Taking the anomaly detection or fault diagnosis task as an example, once data labels are available (e.g., each data samples are associated with their actual condition denoted as normal or faulty), supervised learning algorithms can be readily applied to develop classification models. However, assigning data labels require large amounts of human labor and high level of domain expertise [38]. As a result, the majority of building operational data are unlabeled data, while only a few or even no data samples are properly labeled. The possible drawbacks are two-fold. Firstly, training data-driven models with limited labeled data are prone to underfitting or overfitting problems, which may severely degrade the generalization performance in practice. Secondly, unlabeled building operational data are not compatible with supervised machine learning algorithms, resulting in a large waste of data resources.

As shown in Figure 3, the data imbalance issue may stem from intrinsic building operating characteristics. As an example, anomalies or operating faults are less frequent, indicating that the majority of building operational data corresponds to normal operations and only a few measurements are data samples with user interests. As another example, buildings may have different energy use patterns on weekdays and weekends,

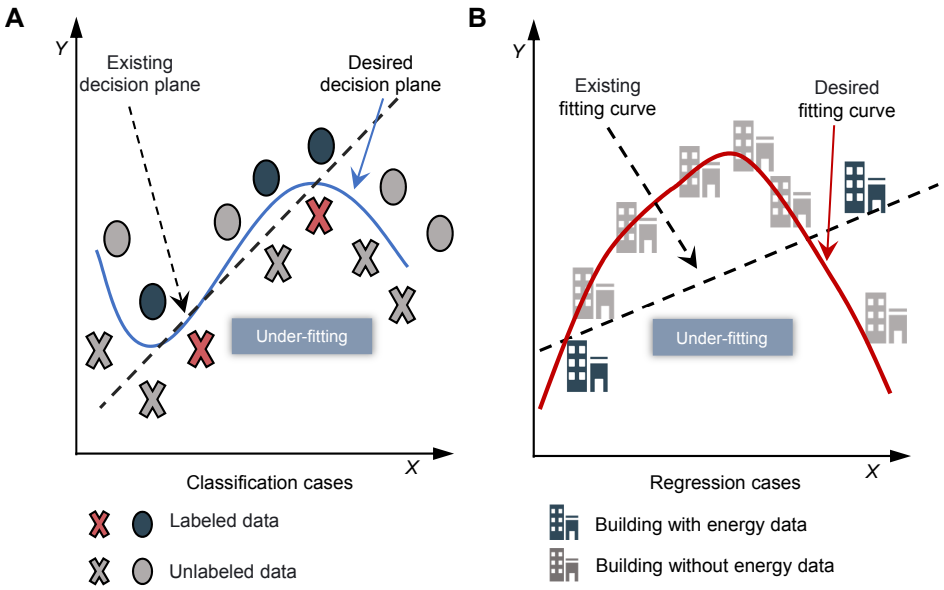

Figure 2 An illustration of poor data-driven model performance due to insufficient labeled data.

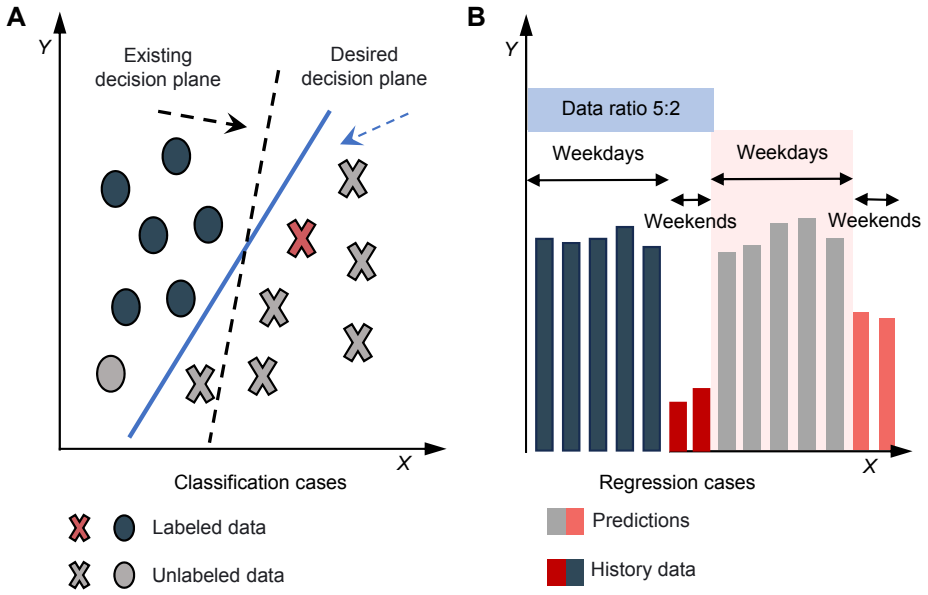

Figure 3 An illustration of poor data-driven model performance due to imbalanced data.

which may also lead to data imbalance problems considering the associated day numbers. Previous studies have also indicated that the numbers and types of buildings used for energy modeling may also suffer from imbalance issues, e.g., most building energy data are collected from commercial buildings, and only a few building samples are available for other types such as museums [39]. Imbalanced data will impose extra challenges for data mining and machine learning algorithms to effectively retrieve patterns from minority data samples [40]. As we shall illustrate in later sessions, data resampling or generative modeling techniques can be used to avoid possible modeling and inference biases caused by imbalanced data.

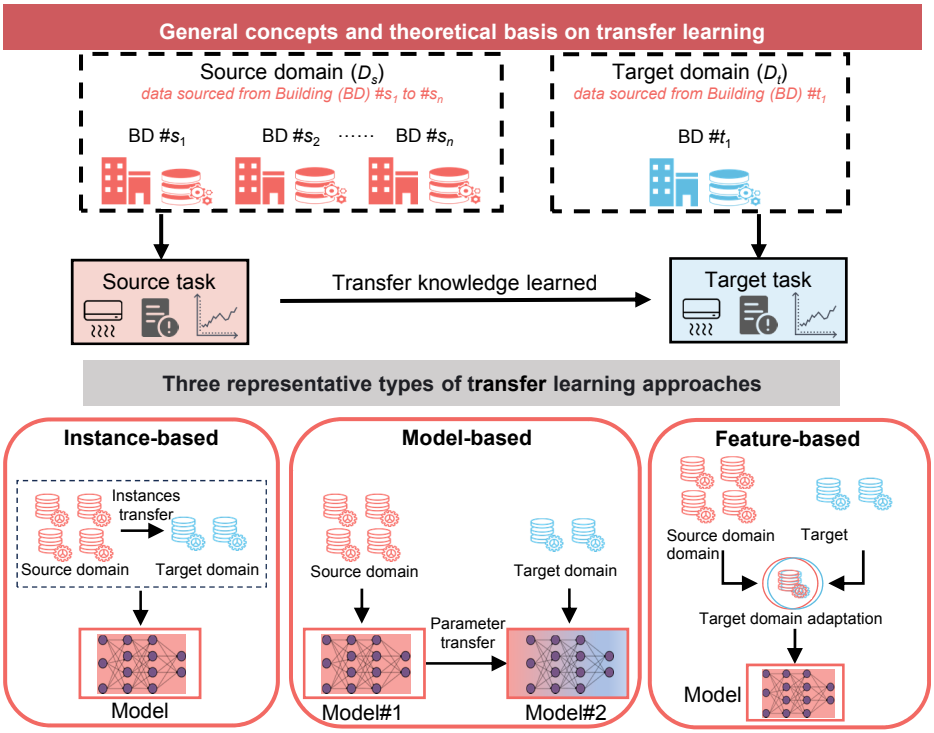

Figure 4 The general concept and representative approaches of transfer learning.

TRANSFER LEARNING-BASED BUILDING DATA ANALYTICS

General concepts and theoretical basis on transfer learning

Transfer learning aims to utilize the knowledge learnt from source domains to facilitate data-driven tasks in target domains [6]. As a formal definition, a domain D is defined as $D= \{X, P(X)\}$ where X and $P(X)$ represent features and their marginal probabilities, respectively. A data-driven task is defined as $T= \{Y, P(Y|X)\}$, where Y and $P(Y|X)$ denote the prediction label and the conditional probability of Y given X, respectively. Transfer learning aims to learn the conditional probability in the target domain $P(Y_t|X_t)$ with the help of knowledge learnt from the source domain D_s and its task T_s, where the subscripts s and t represent source and target domains. As shown in Figure 4, assuming that source and target buildings share similarities in operation patterns, it may be beneficial to utilize the knowledge learnt or operational data collected in the source domain to develop customized prediction models in the target domain. In essence, transfer learning helps to revolutionize the conventional one-to-one relationship between building operational datasets and data-driven models. It has the potential to greatly reduce the training data amounts required and computational costs associated with individual building levels.

There are four general types of transfer learning methods [41]. The first is instance-based method, which incorporates data samples in source domains into the process of developing data-driven models in the target domain through re-weighting or importance sampling techniques [42]. The second is feature-based method, which intends to minimize domain divergence by deriving reliable features from source to target domains [43]. The third is relational knowledge-based method, which focuses on unveiling data relationships in non-

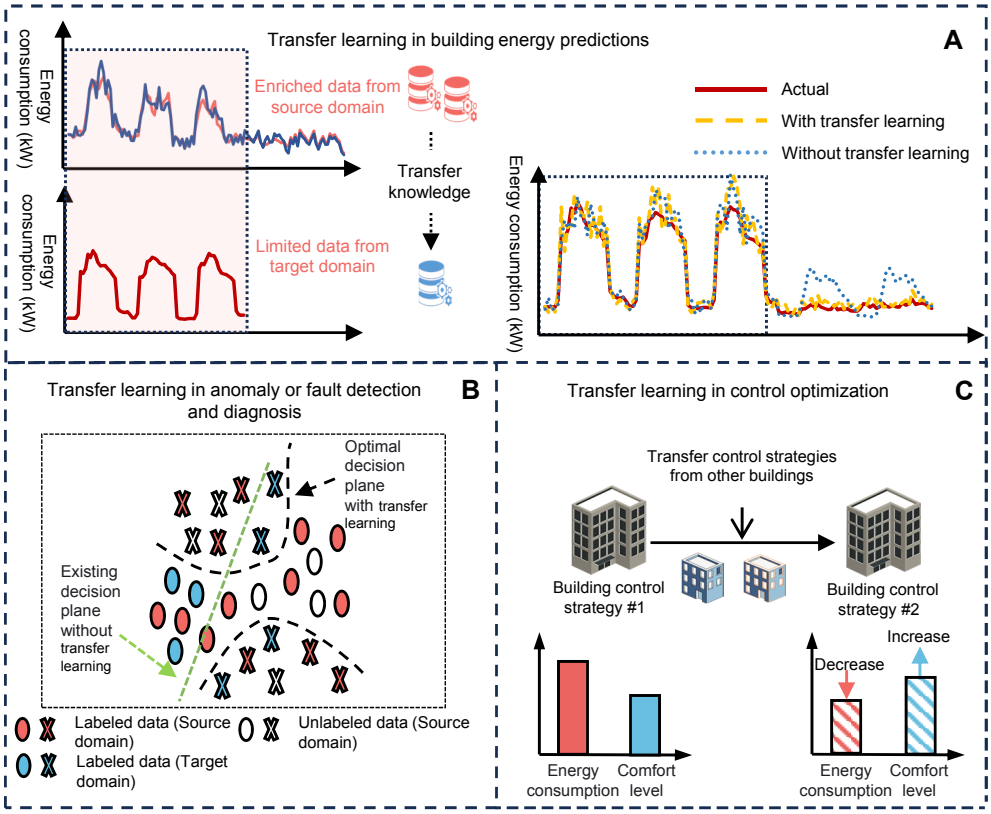

Figure 5 An illustration of transfer learning applications for building energy management.

independent and identically distributed data, e.g., social network data [44]. The fourth is a model parameter-based method, which reuses the model parameters learnt in source domains for prediction tasks in the target domain [45]. Due to the recent success of artificial neural networks and deep learning, the parameter-based methods have gained increasing popularity in knowledge transfer and data sharing. Despite variations in possible neural network-based transfer learning protocols, two main approaches exist using the pre-trained model developed from the source domain for weight initialization and feature extractor, respectively. Existing studies suggested that weight initialization and fine-tuning could bring more consistent transfer learning performance considering unique operating patterns in different buildings [46].

Transfer learning-based building energy management methods

Applications on building energy predictions

As shown in Figure 5, existing studies have explored and validated the value of transfer learning in various tasks. Taking the short-term building energy predictions as an example, Fan *et al*. [47] proposed a transfer learning-based framework to enhance the building energy prediction performance for individual buildings with limited operational data. The operational data from multiple source buildings were used to develop pre-trained, fully connected neural networks, which were then customized for individual buildings through either weight initialization or feature extraction approaches. Data experiments were conducted to simulate two types of data scarcity problems due to large data collection intervals and insufficient data accumulation time,

leading to possible prediction error reductions up to 67%. Rather than using all the available source data, a following study by Fan *et al*. [48] proposed a simple yet effective building similarity metric to select proper source domain data for pre-trained model development. The results indicated that such domain selection process was of great importance to ensure the transfer learning performance, especially when the supervise learning algorithms used are less complicated, e.g., linear regression. To further capture temporal dynamics in building energy data, Lu *et al*. [49] used long short-term memory units to develop prediction models for district energy systems. Customized similarity metrics were proposed to select the most relevant data from source domains, which helps to enhance the transfer learning performance while reducing the chance of negative knowledge transfer. Fang *et al*. [50] proposed a neural network-based method to achieve feature-based transfer learning. The results indicated that by learning time-invariant feature representations between source and target domains, the reliability of prediction models can be further enhanced.

Applications on anomaly or fault detection and diagnosis

As shown in Figure 5B, transfer learning has also been applied to enhance the performance of data-driven anomaly or fault detection and diagnosis methods. The main intuition is to integrate labeled operational data from similar building systems to solve the data-scarce challenge faced by individual buildings. As an example, Liu *et al*. [51] investigated the transfer learning performance in classifying anomalies in building energy systems. The results indicated that useful knowledge learnt from source domains can be effectively migrated to individual buildings through feature extraction and fine-tuning techniques. Zhu *et al*. [52] proposed a feature-based transfer learning method to learnt domain-invariant features for chiller fault diagnosis. Considering that operating conditions between source and target domains may vary, Liang *et al*. [53] developed an adaptive model update method to avoid negative transfer performance. To further address the data compatibility problem among different building systems, Fan *et al*. [46] proposed a novel data transformation method to transform tabular building operational data into 2-dimensional standard image data, which also enables the use of convolutional neural networks for HVAC fault detection and diagnosis tasks. Compared with fully connected neural network, convolutional neural networks are less sensitive to hyper-meters used for updating pre-trained models and thereby, may provide more reliable transfer learning performance.

Applications on control optimization

Extensive studies have been conducted to use model predictive control (MPC) for optimizing building system operations [54]. The key to MPC is to develop accurate and robust prediction models to simulate underlying physical processes. In this regard, transfer learning has been used to ensure the performance of data-driven prediction models in data-scarce contexts. Chen *et al*. [55] adopted multi-layer perceptron models to predict indoor temperature and relative humidity, where the parameter sharing-based transfer learning was used to improve model performance. Transfer learning has also been incorporated with reinforcement learning to achieve reliable optimal controls over building systems. Lissa *et al*. [56] developed a Q-learning method using source domain operational data, which was then migrated to target buildings to achieve efficient room-level thermal controls. Zhang *et al*. [57] combined the use of multi-agent re-

inforcement learning with transfer learning to achieve up to 40% reductions in heating, ventilation and air-conditioning (HVAC) energy consumption. Coraci *et al*. [58] proposed an online transfer learning strategy to migrate deep reinforcement learning-based control strategies learnt from source domains for efficient individual building controls. The results showed that transfer learning was helpful for reducing the time and data required to develop optimal control strategies for HVAC system operations.

To summarize, existing studies mainly used model parameter-based methods to achieve transfer learning between source and target buildings. Fully connected neural networks are mostly used considering their implementation simplicity, while convolutional and recurrent neural networks have also gained successes in addressing the inter-building data compatibility and temporal modeling challenges. Limitations do exist considering possible variations in building functionalities and differences in building occupant behaviors, which requires more stringent methods to ensure domain similarities to avoid negative transfer problems.

SEMI-SUPERVISED LEARNING-BASED BUILDING DATA ANALYTICS

General concepts and theoretical basis on semi-supervised learning

Semi-supervised learning intends to enhance the reliability of data-driven prediction models through effective unlabeled data utilization [7]. It is particularly useful in analyzing building operational data, as the majority of them are unlabeled due to the high costs associated with data labeling. As a formal definition, the data collected can be categorized into three parts. The first is the labeled dataset $\{X_\mathrm{l}, Y_\mathrm{l}\}$ where X and Y denote input and output variables, and the subscript l represents labeled data. The second is the unlabeled dataset $\{X_\mathrm{u}\}$, where the subscript u represents unlabeled data. The third is the testing dataset $\{X_\mathrm{t}\}$, where only the input data X is available, and the subscript t represents testing data. In conventional semi-supervised learning scenarios, the number of labeled data is much less than that of unlabeled data. As shown in Figure 6, the goal of semi-supervised learning is to develop reliable prediction models using limited labeled dataset $\{X_\mathrm{l}, Y_\mathrm{l}\}$ with additional help from the unlabeled dataset $\{X_\mathrm{u}\}$ and thereby, generating accurate predictions for the testing dataset $\{X_\mathrm{t}\}$.

In general, there are three assumptions for semi-supervised learning, i.e., smoothness, cluster and manifold assumptions [7]. The first indicates that two data samples tend to have the same output or labels if they are closely located in a high-density region. The second states that two data samples in the same data cluster are likely to be in the same class or share similar behaviors. The third indicates that if two data samples are close in high dimensional space, then they are also close to each other in low dimensional space. More specifically, there are four main types of semi-supervised learning approaches: generative model-based, low-density separation-based, graph-based and self-labeled approaches [59,60]. In practice, the selection and actual performance of these approaches may vary according to different prediction task formulation and data assumptions. For example, the generative model-based approach assumes that unlabeled data distribution can be described using parametric models, while the low-density separation-based approach assumes decision boundaries only exist in low-density data regions. The graph-based approach typically uses nodes and edges to represent data samples and their similarities, based on which semi-supervised inferences are achieved through graph-based techniques or graph theories, e.g., graph min-cut and label propagation [61,62]. The self-labeled approach works by iteratively using pseudo labels generated from unlabeled data for model

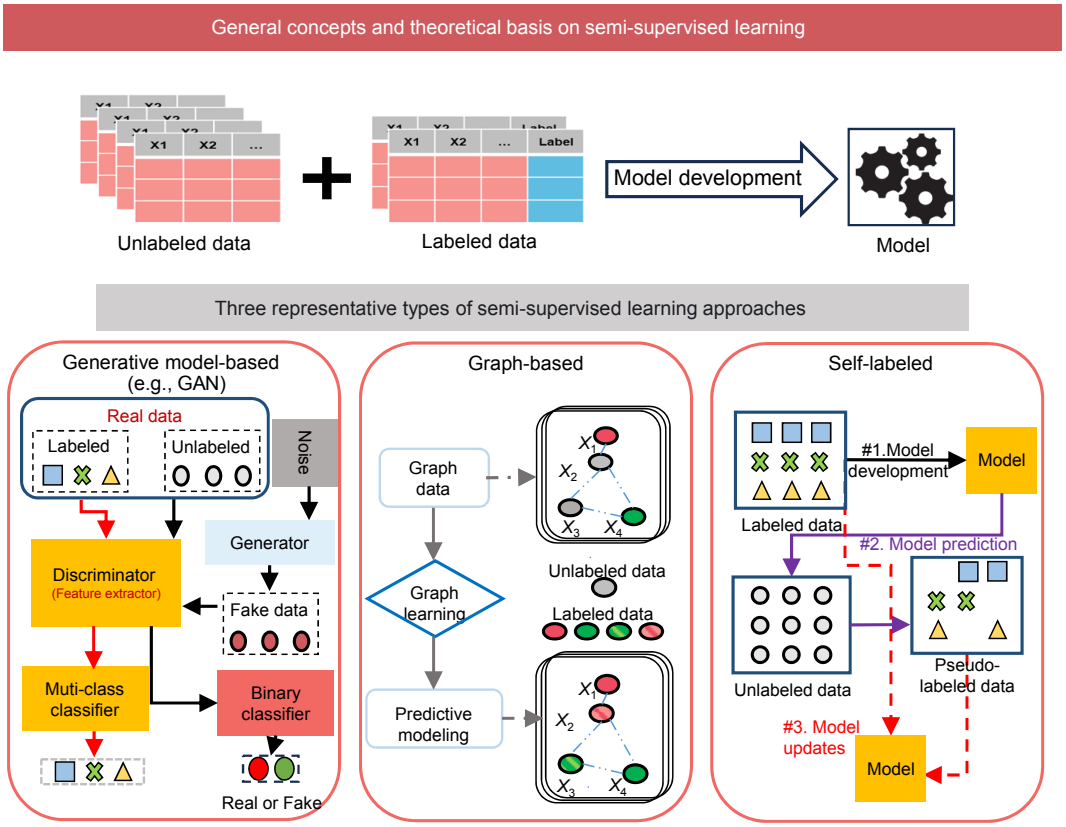

Figure 6 The general concept and representative approaches of semi-supervised learning.

updates and thereby, providing gradual improvements in model quality [63]. In practice, the self-labeled approach may provide more flexibility as it imposes less stringent assumptions on data distributions and can be easily adapted for tabular building operational data [7].

Semi-supervised learning-based building energy management methods

Existing studies have explored the value of semi-supervised learning in both regression and classification tasks for building energy management. As shown in Figure 7A, rather than focusing on energy consumption at individual building levels, existing studies mainly used semi-supervised learning for building energy prediction tasks at district or city levels. The main intuition is that some buildings may only have meta data on building functions, envelope physics and system configurations, while lacking actual energy monitoring data, i.e., making them in essence unlabeled data for energy prediction tasks. In such a case, semi-supervised learning can be incorporated to use information collected from unlabeled buildings to develop more accurate energy prediction models. As an example, Jiang *et al*. [64] developed a deep neural network-based framework to utilize information on buildings without actual energy data for more accurate urban building energy predictions. A dynamic model update scheme was proposed by iteratively choosing the best model trained using both labeled and pseudo-labeled data from k-fold, which in turn helped to generate more reliable pseudo labels and prediction models. Relying on similar ideas, Jin *et al*. [65] proposed a self-labeled method

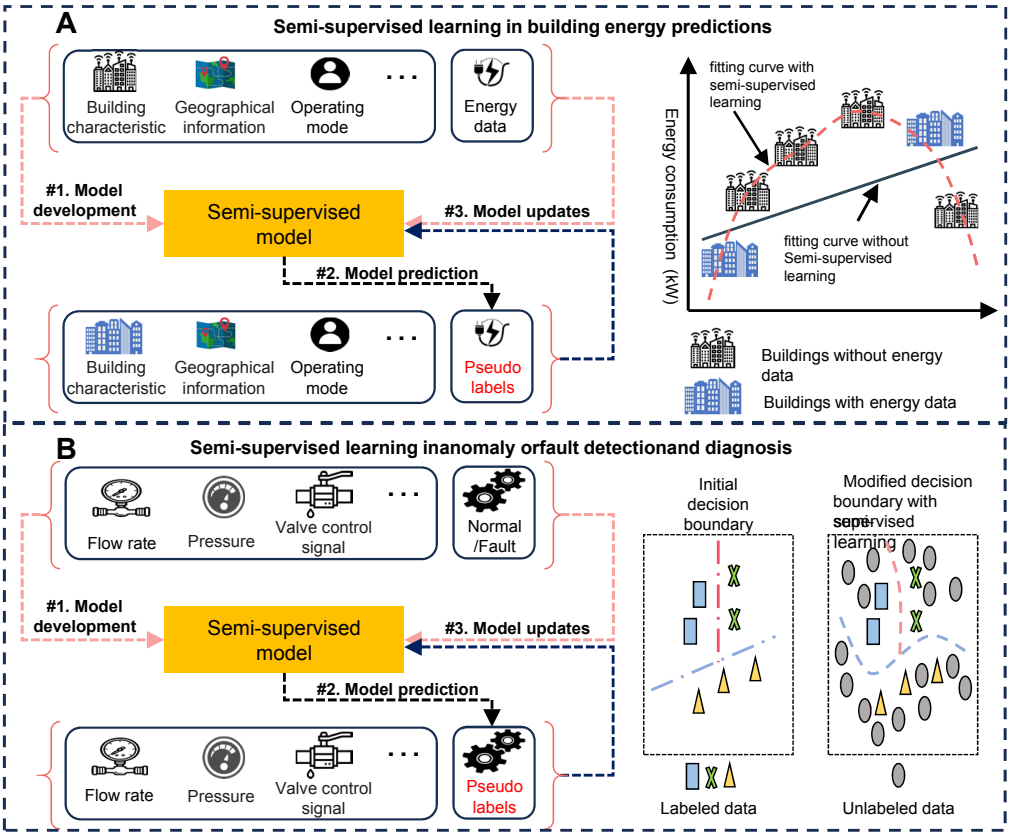

Figure 7 An illustration of semi-supervised learning applications for building energy management.

to predict building energy consumption at district or city levels, where only pseudo energy predictions with relatively high confidence were utilized for updating prediction models. The results showed that up to 10% prediction error reductions could be achieved compared with conventional supervised learning methods.

By contrast, as shown in Figure 7B, emerging studies have mainly used semi-supervised learning to address the anomaly or fault detection and diagnosis tasks, as they can be easily transformed into binary or multi-class classification problems. Tian *et al*. [66] used autoencoders to develop a semi-supervised method for detecting anomalies in building gas systems. The anomalies, such as gas leakage, were successfully identified by comparing the reconstruction errors between normal and abnormal data. Compared with traditional supervised learning methods, semi-supervised learning could help to achieve satisfactory anomaly detection performance with much less labeled data. Yan *et al*. [67] compared the performance of several off-the-shelf semi-supervised learning algorithms, such as semi-supervised support vector machines, in analyzing air handling unit (AHU) operating faults. Fan *et al*. [68] developed a semi-supervised neural network-based methods for classifying faults in AHUs, where the self-labeled approach was adopted to generate high-quality pseudo labeled data for model updates. The results showed that semi-supervised learning could bring evident enhancements in classifying both seen and unseen faults in the labeled data. Semi-supervised learning has also been coupled with advanced generative learning techniques to enhance the HVAC fault classification problem. As an example, Li *et al*. [69] developed generative adversarial networks to achieve semi-supervised fault diagnosis for chillers. More specifically, the discriminator was trained to classify actual

labels of real data while distinguishing between real and synthetic data samples, leading to more reliable fault diagnosis performance with limited labeled data. Existing studies have also explored the value of graph-based semi-supervised learning for HVAC fault classification tasks [70]. The main idea was to treat each data sample, either labeled or unlabeled, as a node and create edges based on inter-node similarities to allow possible label inferences or label propagation using graph convolutional neural networks. Despite the type of graph convolutional networks used, evident boosts in fault classification accuracies were obtained across three different AHU operational datasets.

To summarize, semi-supervised learning has been mainly used to handle classification problems in the building field. Various semi-supervised learning methods have been used, ranging from off-the-shelf algorithms to customized frameworks using self-labeled or generative modeling techniques. Self-labeled methods have gained great popularity due to their implementation simplicity and compatibility with artificial neural networks (e.g., SoftMax activation can be used at the output layer to generate scores for evaluating pseudo label quality). However, as reported by existing studies, such methods may fail when the initial labeled data are not sufficient to generate reliable and generalizable results [63,68]. In such a case, the pseudo labels generated from unlabeled data may be of poor quality, leading to negative impacts on model updates.

GENERATIVE LEARNING-BASED BUILDING DATA ANALYTICS

General concepts and theoretical basis on generative learning

Generative learning aims to capture the distribution of real measurements and thereby, enabling the generation of high-quality synthetic data samples with physical meanings. It has proven to be useful to solve data shortage or imbalance problems in various fields [71].

As shown in Figure 8, the recent advances in deep learning have provided powerful tools for generative learning and two representatives are generative adversarial networks (i.e., GANs) [72] and variational autoencoders (i.e., VAEs) [73]. Such tools have shown excellent abilities in creating high-quality synthetic data such as images and time-series [74,75]. A GAN model consists of a generator and a discriminator model. In basic settings, the generator model takes random noises as inputs to generate synthetic data, while the discriminator model is trained to classify whether a data sample is real and fake. In essence, GANs are trained to achieve the Nash equilibrium of a zero-sum game between discriminator and generator, i.e., the generator tries to generate synthetic data to fool the discriminator, while the discriminator is trained with the opposite objective. VAEs are variants of autoencoders, which consist of an encoder and a decoder model for data generation [76]. In basic settings, the VAE encoder model will transform the original data into a number of means and variances to specify normal distribution characteristics. Random sampling is performed to draw a random vector from such distributions to form a latent vector, which is then fed to the VAE decoder model to generate synthetic data samples. VAE models are typically trained to minimize two types of losses, i.e., one for evaluating data reconstruction error between original and synthetic data, and one regularization loss for avoiding overfitting.

In practice, GAN models are highly sensitive to hyperparameters (e.g., learning rates) and model architectures. Extensive tunings are expected to ensure the quality of synthetic data. By contrast, VAEs are much easier to train using gradient-based methods [73]. Existing studies in the computer science suggested that

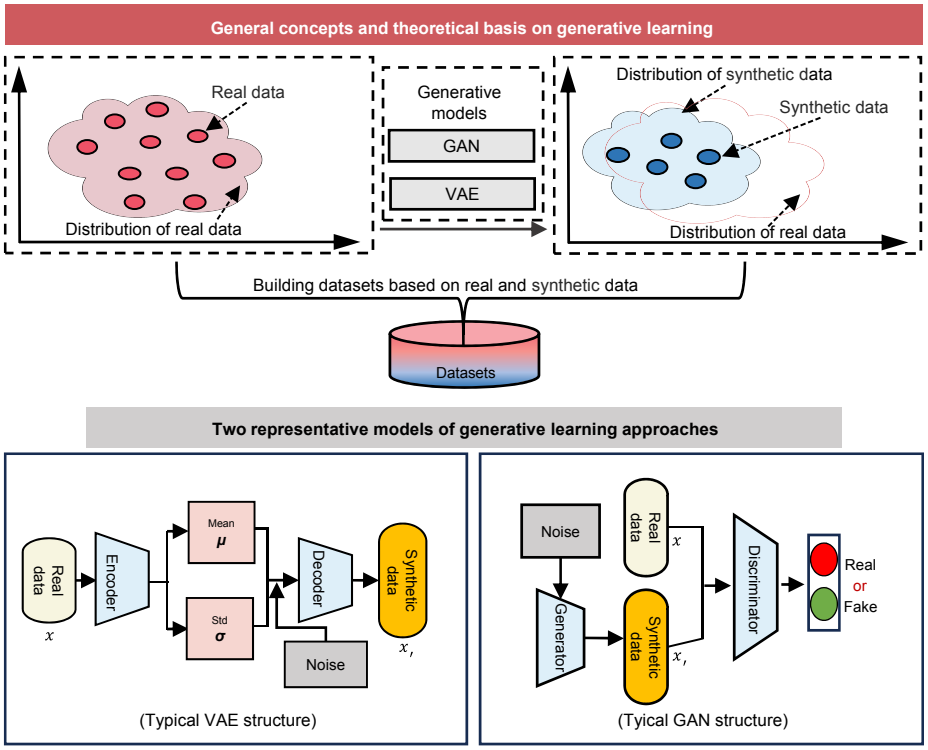

Figure 8 The general concept and representative approaches of generative learning.

GANs could generate more authentic data in various industries [77]. However, considering that building data are typically tabular data with relatively low complexity, VAEs may be more suitable for practice due to their ease of implementations.

Generative learning-based building energy management methods

Applications on building energy predictions

As illustrated in Figure 9A, synthetic building energy data have proven to be useful in enhancing accuracies of various data-driven energy management tasks. Tian *et al*. [78] used GANs to create synthetic daily building energy data to enhance the performance of building energy consumption, where fully connected and recurrent neural networks were developed as the generator and discriminator model, respectively. The results obtained verified that synthetic energy data could lead to better performance compared with conventional methods such as information diffusion and bootstrap resampling. To capture temporal dependencies in building operations, Zhang *et al*. [79] proposed a GAN model suitable for generating synthetic time-series data. The method was applied to learn data distributions in different heating periods and effectively enhanced the reliability of heating load predictions. GAN models are typically of high complexity, and it can be very challenging to find the optimal and balancing configures for generator and discriminator models. To tackle such challenges, Fan *et al*. [80] used conditional VAEs to generate weekly energy data considering different seasons and day types. The potential of synthetic data and different VAE architectures was verified using 24-hour ahead prediction tasks on over 50 buildings, resulting in averaged error reduction ratios of 12%–18%.

To ensure the quality of synthetic energy data, it is often required to have sufficient data for developing generative models. Given limited operational data, Heidrich *et al*. [81] proposed an integrated strategy to combine transfer learning and generative learning for data generation. Transfer learning was first applied to learn and match the latent data space between source and target domains, based on which synthetic data were then generated for training building energy prediction models.

Applications on anomaly or fault detection and diagnosis

As illustrated in Figure 9B, synthetic operational data have proven to be useful in building-related classification tasks. Zhong *et al*. [82] used Wasserstein GANs to generate synthetic AHU operational data, which helped to achieve over 90% fault diagnosis accuracies given limited labeled data. Yan *et al*. [83] developed conditional GANs to generate synthetic data for different faulty conditions, which helped to address the potential data imbalance issue in AHU fault diagnosis. Similarly, GAN models have been used to generate synthetic chiller operation data, resulting in more reliable data-driven fault classification models [84]. Fan *et al*. [85] conducted comprehensive data experiments to compare the value of different data augmentation methods in HVAC fault diagnosis tasks. A wide range of data augmentation methods, ranging from conditional VAEs to conventional random over- or under-sampling and synthetic minority over-sampling techniques, were used to generate synthetic data for AHU fault classification tasks. The results showed that the synthetic data generator was particularly useful in imbalanced data-scarce contexts, which could lead to an average accuracy increase of up to 7.92%. The research results indicated that deep generative learning methods, such as VAEs, could be coupled with conventional data augmentation techniques to further enhance the quality of synthetic operational data. Similarly, Zhang *et al*. [86] integrated GANs and VAEs to form an end-to-end training scheme for synthetic data generation, which led to evident improvements in VRF fault diagnosis tasks.

Applications on control optimizations

As illustrated in Figure 9C, An emerging application of generative learning is to generate possible yet unseen operating scenarios to enhance the generalizability of control strategies. Li *et al*. [87] adopted GAN models to generate multiple operating conditions for emergency management, which helped to enhance the performance in assessing short-term electricity voltage stability. Wang *et al*. [88] used Wasserstein GANs to simulate variability in renewable energy generations, enabling the development of more efficient control strategies for economic and reliable operations. Lin *et al*. [89] developed a GAN-based method to simulate occupant behaviors in building demand responses, which in turn helped to optimize electricity prices with the help of reinforcement learning.

To summarize, generative learning using neural network-based methods has proven to be useful in building energy predictions, anomaly or fault detection and diagnosis, and control optimization tasks. Once trained properly, generative models have the ability to generate synthetic yet meaningful operational data. Such data can be used to tackle data imbalance issues (e.g., generating more data on faulty conditions to reduce inference biases in data-driven classification models), simulate possible unseen operation scenarios for enhancing the generalizability of control strategies, or simply serve as a data-preprocessing step to meet the

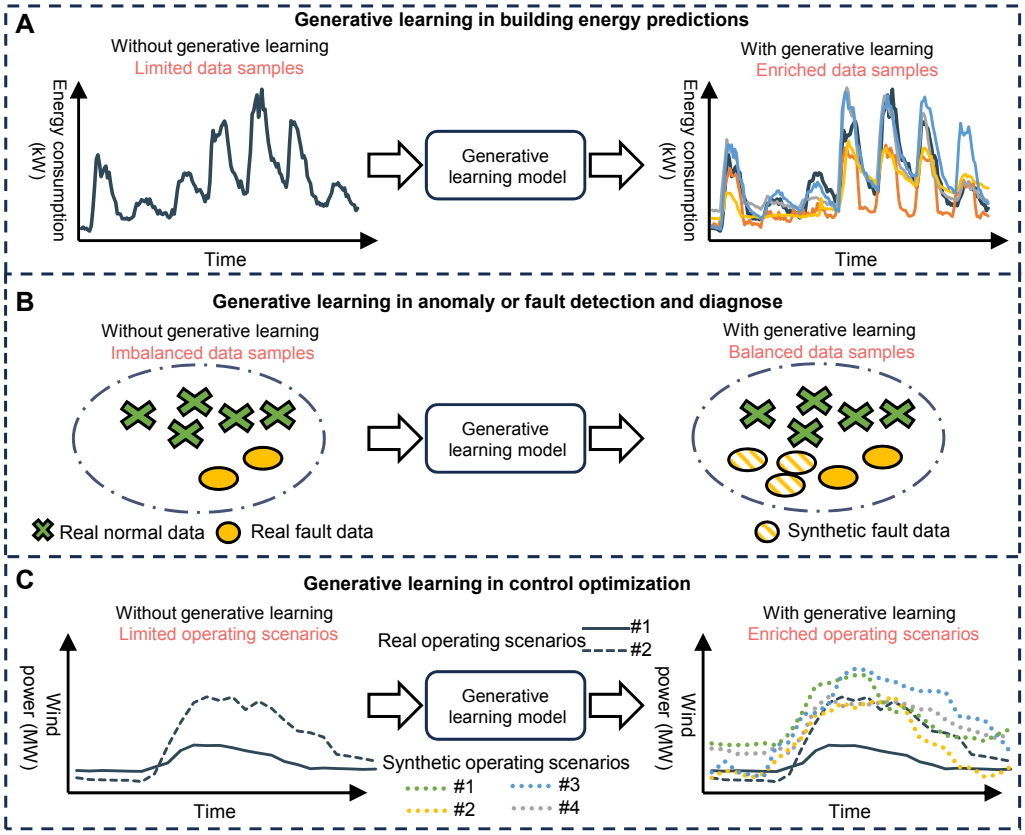

Figure 9 An illustration of generative learning applications for building energy management.

requirements of using complicated machine learning algorithms. Nevertheless, it should be noted that generative models (e.g., GANs) can be difficult to develop as they are sensitive to training hyperparameters and protocols. In addition, some generative models have intrinsic assumptions, e.g., the underlying data follow a certain distribution. If violated, the synthetic data generated may result in negative impacts on predictive modeling, e.g., generating synthetic data with normal distributions may not be helpful if actual measurements follow bimodal distributions. Such challenges also call for more systematic and stringent methods to evaluate the similarity between synthetic and generated data for valid applications.

FUTURE PERSPECTIVES

Development of generalizable implementation guidelines for practical applications

As reviewed above, novel machine learning paradigms, such as transfer learning, semi-supervised learning and generative learning, have been successfully used to tackle data challenges in building operations. Despite their promising potential, the underlying theories are quite complicated for building professionals and their performance can be highly sensitive to implementation protocols. Taking transfer learning as an example, there are two key steps to ensure its validity and avoid negative transfer, i.e., source domain selection and knowledge transfer methods. Both steps require interdisciplinary expertise and therefore, implementation

references or guidelines are needed to ensure its applicability in practice. Similarly, the application of deep generative models can be quite challenging due to the difficulties in training complex GAN models. Existing studies in computer science mainly focus on generating unstructured data such as images or texts, which provides little references for tabular building operational data. Little to none consensus has been reached in generating high-quality building operational data with tabular formats, e.g., which model architecture is the most suitable one for generating synthetic yet meaningful energy or system operation data. In this regard, large-scale data experiments are required to derive generalizable implementation guidelines for efficient and effective applications in the building field.

Other solutions for the efficient utilization of unlabeled building operational data

Most building operational data are collected with an unlabeled nature. Other machine learning paradigms do exist and may help to exploit the value in massive amounts of unlabeled building operational data. Taking self-supervised learning as an example, it provides an alternative approach to learning from unlabeled data through user-defined pre-text tasks, e.g., artificially rotating images to different degrees to form multi-class classification pre-text tasks [90]. Considering the tabular nature of building operational data, research effort is needed to design suitable pre-text tasks to ensure knowledge discovery efficiency. An exploratory study has been conducted to use data reconstruction and noise location identification methods to enable self-supervised learning from unlabeled building operational data [91]. Promising results have been obtained as it helped to greatly reduce data labeling costs while achieving similar performance on several HVAC fault diagnosis tasks. Further studies may focus on designing suitable pre-text tasks to extract temporal dynamics in building operational data while enlarging experiment datasets to draw more generalizable results.

Besides purely data-centric solutions, it is observed that human involvement may become inevitable in extreme data-scarce contexts. In such a case, it is desired to develop strategies to maximize the cost-effectiveness of human labor. Active learning, which is an iterative learning paradigm aiming to produce high-quality data-driven models by smartly choosing informative unlabeled data for manual labeling, may become useful for building professionals [92]. It can be of great value in enhancing the efficiency of building professionals. However, at present, few studies have been conducted to explore the value of active learning in the building, and the main application is to select simulation settings or operating conditions to minimize uncertainties in building energy simulation [93] or fault diagnosis tasks [94]. Further studies are needed to provide efficient protocols or methods to integrate domain expertise with massive data resources for smart decision makings.

Promoting integration at both data and learning paradigm levels

Smart building operations require support from all possible information sources. Besides tabular operational data collected by building automation systems, other data, such as video surveillance data and textual maintenance data, are also available for possible knowledge discovery. The integration at data level encourages the unification of multi-source information, which in turn requires the development of suitable data analytics capable of addressing the multi-model data analysis challenge. As an example, Jacoby *et al*. [95] proposed a sensor fusion framework to integrate image data on indoor spaces, time-series data on indoor

audio and environmental conditions for building occupancy detection. Given sufficient considerations on human privacy, further studies may follow similar ideas to design data analytics compatible with multi-modality data, such as occupant behavior data collected through cameras or social media, textual data in system or component maintenance reports, building structural data from oscillation or displacement sensors, etc.

Besides data-level integration, it is also promising to incorporate multiple machine learning paradigms to effectively address data challenges in the building field. For instance, semi-supervised learning and active learning can be readily integrated to address the challenge of insufficient labeled data. Taking the multi-class fault classification task as an example, some unlabeled data are farther away from decision boundaries of existing data-driven models, and their pseudo labels generated from semi-supervised learning can be used as confident evidence for model updates. Meanwhile, active learning can be used to find informative data samples in unlabeled data (e.g., those lying closely with existing model decision boundaries) for manual labeling. As a result, the integration of these two learning paradigms can provide efficient and effective streamlines to ensure fault classification performance at minimal labeled data and human laboring costs.

CONCLUSIONS

Despite the wide adoption of information technologies in buildings, various data challenges do exist, which hinder the applications of data-driven technologies for smart building operations. This review summarizes three key data challenges for building system predictive modeling, i.e., the lack of data representativeness, insufficient labeled data, and imbalanced data, while presenting the recent progress using novel machine learning paradigms as possible solutions. Besides theoretical basis, representative studies has been reviewed focusing on typical applications in building energy management, including energy prediction, anomaly or fault detection and diagnosis, and control optimization. Existing studies have shown that transfer learning can help solve the data shortage problem in individual buildings by leveraging data resources from similar buildings. Semi-supervised learning provides a solution to integrate massive amounts of unlabeled operational data for predictive modeling, while generative modeling can generate valuable synthetic data to tackle possible data imbalance and limited data variability problems. To further facilitate the upgrades in building data analytics, future perspectives have been provided on three main aspects, i.e., deriving easy-to-use and generalizable implementation guidelines, exploring other machine learning paradigms to effectively utilize unlabeled building operational data with minimal human interventions, and developing integrated solutions at both data and learning paradigm levels.

Data availability

The original data are available from corresponding authors upon reasonable request.

Funding

This work was supported by the National Natural Science Foundation of China (52278117), the Philosophical and Social Science Program of Guangdong Province, China (GD22XGL20), and the Shenzhen Science and Technology Program (20220531101800001 and 20220810160221001).

Author contributions

C.F. conceived the concept, designed the methodology, and prepared the original draft. Y.L. conducted the investigation and prepared the visualization. J.M. and H.W. contributed to the methodology and investigation. Q.W. and J.C. participated in the investigation and visualization.

Conflict of interest

The authors declare no conflict of interest.

References

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

2 Chen Z, O'Neill Z, Wen J, *et al.* A review of data-driven fault detection and diagnostics for building HVAC systems. *Appl Energy* 2023; **339**: 121030.

3 Dong B, Prakash V, Feng F, *et al.* A review of smart building sensing system for better indoor environment control. *Energy Build* 2019; **199**: 29–46.

4 Halhoul Merabet G, Essaaidi M, Ben Haddou M, *et al.* Intelligent building control systems for thermal comfort and energy-efficiency: A systematic review of artificial intelligence-assisted techniques. *Renew Sustain Energy Rev* 2021; **144**: 110969.

5 Chen Y, Xu P, Gu J, *et al.* Measures to improve energy demand flexibility in buildings for demand response (DR): A review. *Energy Build* 2018; **177**: 125–139.

6 Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. *J Big Data* 2016; **3**: 1–40.

7 van Engelen JE, Hoos HH. A survey on semi-supervised learning. *Mach Learn* 2020; **109**: 373–440.

8 Gm H, Gourisaria MK, Pandey M, *et al.* A comprehensive survey and analysis of generative models in machine learning. *Comput Sci Rev* 2020; **38**: 100285.

9 Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. *J Artif Intell Res* 1996; **4**: 237–285.

10 Ye R, Dai Q. Implementing transfer learning across different datasets for time series forecasting. *Pattern Recogn* 2021; **109**: 107617.

11 Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal* 2019; **54**: 280–296.

12 Nagy Z, Henze G, Dey S, *et al.* Ten questions concerning reinforcement learning for building energy management. *Build Environ* 2023; **241**: 110435.

13 Weinberg D, Wang Q, Timoudas TO, *et al.* A review of reinforcement learning for controlling building energy systems from a computer science perspective. *Sustain Cities Soc* 2023; **89**: 104351.

14 Chen Z, Xiao F, Guo F, *et al*. Interpretable machine learning for building energy management: A state-of-the-art review *Advances Appl Energy* 2023; **9**:100123.

15 Zhang C, Tian X, Zhao Y, *et al.* Automated machine learning-based building energy load prediction method. *J Build Eng* 2023; **80**: 108071.

16 Mao Y, Yu J, Zhang N, *et al.* A hybrid model of commercial building cooling load prediction based on the improved NCHHO-FENN algorithm. *J Build Eng* 2023; **78**: 107660.

17 Lu C, Li S, Reddy Penaka S, *et al.* Automated machine learning-based framework of heating and cooling load prediction for quick residential building design. *Energy* 2023; **274**: 127334.

18 Wang X, Yao Z, Papaefthymiou M. A real-time electrical load forecasting and unsupervised anomaly detection framework. *Appl Energy* 2023; **330**: 120279.

19 Fan C, Xiao F, Zhao Y, *et al.* Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data. *Appl Energy* 2018; **211**: 1123–1135.

20 Blum D, Wang Z, Weyandt C, *et al.* Field demonstration and implementation analysis of model predictive control in an office HVAC system. *Appl Energy* 2022; **318**: 119104.

21 Sepehri A, Pavlak GS. Evaluating optimal control of active insulation and HVAC systems in residential buildings. *Energy Build* 2023; **281**: 112728.

22 Zhao T, Zhang B, Li M, *et al.* Handling fault detection and diagnosis in incomplete sensor measurements for BAS based HVAC system. *J Build Eng* 2023; **80**: 108098.

23 Chen Z, Xiao F, Guo F. Similarity learning-based fault detection and diagnosis in building HVAC systems with limited labeled data. *Renew Sustain Energy Rev* 2023; **185**: 113612.

24 Heimar Andersen K, Pommerencke Melgaard S, Johra H, *et al.* Barriers and drivers for implementation of automatic fault detection and diagnosis in buildings and HVAC systems: An outlook from industry experts. *Energy Build* 2024; **303**: 113801.

25 Wen S, Zhang W, Sun Y, *et al.* An enhanced principal component analysis method with Savitzky-Golay filter and clustering algorithm for sensor fault detection and diagnosis. *Appl Energy* 2023; **337**: 120862.

26 Gunay HB, Shi Z. Cluster analysis-based anomaly detection in building automation systems. *Energy Build* 2020; **228**: 110445.

27 Yang XB, Jin XQ, Du ZM, *et al.* A hybrid model-based fault detection strategy for air handling unit sensors. *Energy Build* 2013; **57**: 132–143.

28 Liang A, Hu Y, Li G. The impact of improved PCA method based on anomaly detection on chiller sensor fault detection. *Int J Refriger* 2023; **155**: 184–194.

29 Cheng F, Cai W, Zhang X, *et al.* Fault detection and diagnosis for Air Handling Unit based on multiscale convolutional neural networks. *Energy Build* 2021; **236**: 110795.

30 Li G, Chen L, Fan C, *et al.* Interpretation and explanation of convolutional neural network-based fault diagnosis model at the feature-level for building energy systems. *Energy Build* 2023; **295**: 113326.

31 Wang X, Kang X, An J, *et al.* Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. *Energy Build* 2023; **301**: 113696.

32 Georgiou GS, Christodoulides P, Kalogirou SA. Optimizing the energy storage schedule of a battery in a PV grid-connected nZEB using linear programming. *Energy* 2020; **208**: 118177.

33 Čongradac V, Kulić F. Recognition of the importance of using artificial neural networks and genetic algorithms to optimize chiller operation. *Energy Build* 2012; **47**: 651–658.

34 Delgarm N, Sajadi B, Kowsary F, *et al.* Multi-objective optimization of the building energy performance: A simulation-based approach by means of particle swarm optimization (PSO). *Appl Energy* 2016; **170**: 293–303.

35 Chan KC, Wong VTT, Yow AKF, *et al.* Development and performance evaluation of a chiller plant predictive operational control strategy by artificial intelligence. *Energy Build* 2022; **262**: 112017.

36 Qiu S, Li Z, Fan D, *et al.* Chilled water temperature resetting using model-free reinforcement learning: Engineering application. *Energy Build* 2022; **255**: 111694.

37 Genkin M, McArthur J. A transfer learning approach to minimize reinforcement learning risks in energy optimization for automated and smart buildings. *Energy Build* 2024; **303**: 113760.

38 Zhang Q, Tian Z, Niu J, *et al.* A study on transfer learning in enhancing performance of building energy system fault diagnosis with extremely limited labeled data. *Build Environ* 2022; **225**: 109641.

39 Jin X, Xiao F, Zhang C, *et al*. GEIN: An interpretable benchmarking framework towards all building types based on machine learning. *Appl Energy* 2022; **260**: 111909.

40 Zhong F, Calautit JK, Wu Y. Fault data seasonal imbalance and insufficiency impacts on data-driven heating, ventilation and air-conditioning fault detection and diagnosis performances for energy-efficient building operations. *Energy* 2023; **282**: 128180.

41 Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2009; **22**: 1345–1359.

42 Wang T, Huan J, Zhu M. Instance-based deep transfer learning. In: Proceedings of the 2019 IEEE Winter Conference on

Applications of Computer Vision. Waikoloa, 2019. 367–375.

43 Niu S, Hu Y, Wang J, *et al*. Feature-based distant domain transfer learning. In: Proceedings of the Programs and Abstracts of the 2020 IEEE International Conference on Big Data. Atlanta, 2020. 5164–5171.

44 Blome C, Schoenherr T, Eckstein D. The impact of knowledge transfer and complexity on supply chain flexibility: A knowledge-based view. *Int J Prod Econ* 2014; **147**: 307–316.

45 Pinto G, Messina R, Li H, *et al.* Sharing is caring: An extensive analysis of parameter-based transfer learning for the prediction of building thermal dynamics. *Energy Build* 2022; **276**: 112530.

46 Fan C, He W, Liu Y, *et al.* A novel image-based transfer learning framework for cross-domain HVAC fault diagnosis: From multi-source data integration to knowledge sharing strategies. *Energy Build* 2022; **262**: 111995.

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

48 Fan C, Lei Y, Sun Y, *et al.* Data-centric or algorithm-centric: Exploiting the performance of transfer learning for improving building energy predictions in data-scarce context. *Energy* 2022; **240**: 122775.

49 Lu Y, Tian Z, Zhou R, *et al.* A general transfer learning-based framework for thermal load prediction in regional energy system. *Energy* 2021; **217**: 119322.

50 Fang X, Gong G, Li G, *et al.* A hybrid deep transfer learning strategy for short term cross-building energy prediction. *Energy* 2021; **215**: 119208.

51 Liu J, Zhang Q, Li X, *et al.* Transfer learning-based strategies for fault diagnosis in building energy systems. *Energy Build* 2021; **250**: 111256.

52 Zhu X, Chen K, Anduv B, *et al.* Transfer learning based methodology for migration and application of fault detection and diagnosis between building chillers for improving energy efficiency. *Build Environ* 2021; **200**: 107957.

53 Liang X, Li P, Chen S, *et al.* Partial domain adaption based prediction calibration methodology for fault detection and diagnosis of chillers under variable operational condition scenarios. *Build Environ* 2022; **217**: 109099.

54 Shaikh PH, Nor NBM, Nallagownden P, *et al.* A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renew Sustain Energy Rev* 2014; **34**: 409–429.

55 Chen Y, Tong Z, Zheng Y, *et al.* Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings. *J Clean Prod* 2020; **254**: 119866.

56 Lissa P, Schukat M, Barrett E. Transfer learning applied to reinforcement learning-based HVAC control. *SN Comput Sci* 2020; **1**: 127.

57 Zhang T, Afshari M, Musilek P, *et al*. Diversity for transfer in learning-based control of buildings. In: Proceedings of the Thirteenth ACM International Conference on Future Energy Systems. Association for Computing Machinery. New York, 2022. 556–564.

58 Coraci D, Brandi S, Hong T, *et al.* Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. *Appl Energy* 2023; **333**: 120598.

59 Chong Y, Ding Y, Yan Q, *et al.* Graph-based semi-supervised learning: A review. *Neurocomputing* 2020; **408**: 216–230.

60 Triguero I, García S, Herrera F. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowl Inf Syst* 2015; **42**: 245–284.

61 Karger DR, Stein C. A new approach to the minimum cut problem. *J ACM* 1996; **43**: 601–640.

62 Zang F, Zhang JS. Label propagation through sparse neighborhood and its applications. *Neurocomputing* 2012; **97**: 267–277.

63 Fan C, Liu Y, Liu X, *et al.* A study on semi-supervised learning in enhancing performance of AHU unseen fault detection with limited labeled data. *Sustain Cities Soc* 2021; **70**: 102874.

64 Jiang F, Ma J, Li Z, *et al.* Prediction of energy use intensity of urban buildings using the semi-supervised deep learning model. *Energy* 2022; **249**: 123631.

65 Jin X, Xiao F, Zhang C, *et al.* Semi-supervised learning based framework for urban level building electricity consumption prediction. *Appl Energy* 2022; **328**: 120210.

66 Tian X, Jiao W, Liu T, *et al.* Intelligent detection method of low-pressure gas system leakage based on semi-supervised anomaly diagnosis. *Expert Syst Appl* 2022; **209**: 118376.

67 Yan K, Zhong C, Ji Z, *et al.* Semi-supervised learning for early detection and diagnosis of various air handling unit faults. *Energy Build* 2018; **181**: 75–83.

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

69 Li B, Cheng F, Cai H, *et al.* A semi-supervised approach to fault detection and diagnosis for building HVAC systems based on the modified generative adversarial network. *Energy Build* 2021; **246**: 111044.

70 Fan C, Lin Y, Piscitelli MS, *et al*. Leveraging graph convolutional networks for semi-supervised fault diagnosis of HVAC systems in data-scarce contexts. *Build Simul* 2023; **16**: 1499–1517.

71 Bond-Taylor S, Leach A, Long Y, *et al*. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. In: Proceedings of the Programs and Abstracts of the IEEE Transactions on Pattern Analysis and Machine Intelligence. Institute of Electrical and Electronics Engineers, Piscataway, 2021. 7327–7347.

72 Creswell A, White T, Dumoulin V, *et al.* Generative adversarial networks: An overview. *IEEE Signal Proc Mag* 2018; **35**: 53–65.

73 Kingma DP, Welling M. An introduction to variational autoencoders. *FNT Machine Learn* 2019; **12**: 307–392.

74 Abufadda M, Mansour K. A survey of synthetic data generation for machine learning. In: Proceedings of the Programs and Abstracts of the 22nd International Arab Conference on Information Technology. Muscat, 2021.

75 Debnath A, Waghmare G, Wadhwa H, *et al*. Exploring generative data augmentation in multivariate time series forecasting: Opportunities and challenges. In: Proceedings of the 7th KDD Workshop on Mining and Learning from Time Series. Singapore, 2021.

76 Zhai J, Zhang S, Chen J, *et al.* Autoencoder and its various variants. In: Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics. Miyazaki, 2018. 415–419.

77 Gui J, Sun Z, Wen Y, *et al*. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans Pattern Anal Mach Intell* 2021; **35**: 3313–3332.

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

79 Zhang Y, Zhou Z, Liu J, *et al.* Data augmentation for improving heating load prediction of heating substation based on TimeGAN. *Energy* 2022; **260**: 124919.

80 Fan C, Chen M, Tang R, *et al.* A novel deep generative modeling-based data augmentation strategy for improving short-term building energy predictions. *Build Simul* 2022; **15**: 197–211.

81 Heidrich B, Mannsperger L, Turowski M, *et al.* Boost short-term load forecasts with synthetic data from transferred latent space information. *Energy Inform* 2022; **5**: 1–20.

82 Zhong C, Yan K, Dai Y, *et al.* Energy efficiency solutions for buildings: Automated fault diagnosis of air handling units using generative adversarial networks. *Energies* 2019; **12**: 527.

83 Yan K, Huang J, Shen W, *et al.* Unsupervised learning for fault detection and diagnosis of air handling units. *Energy Build* 2020; **210**: 109689.

84 Yan K, Chong A, Mo Y. Generative adversarial network for fault detection diagnosis of chillers. *Build Environ* 2020; **172**: 106698.

85 Fan C, Li X, Zhao Y, *et al.* Quantitative assessments on advanced data synthesis strategies for enhancing imbalanced AHU fault diagnosis performance. *Energy Build* 2021; **252**: 111423.

86 Zhang F, Saeed N, Sadeghian P. Deep learning in fault detection and diagnosis of building HVAC systems: A systematic review with meta analysis. *Energy AI* 2023; **12**: 100235.

87 Li Y, Zhang M, Chen C. A Deep-Learning intelligent system incorporating data augmentation for Short-Term voltage stability assessment of power systems. *Appl Energy* 2022; **308**: 118347.

88 Wang Y, Liu Y, Yang Q. Operational scenario generation and forecasting for integrated energy systems. *IEEE Trans Indust Inf* 2023; **20**: 2920–2931.

89 Lin J, Zhang Y, Xu S. Improved generative adversarial behavioral learning method for demand response and its application in hourly electricity price optimization. *J Modern Power Syst Clean Energy* 2022; **10**: 1358–1373.

90 Feng Z, Xu C, Tao D. Self-supervised representation learning from multi-domain data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, 2019. 3245–3255.

91 Fan C, Lei Y, Sun Y, *et al.* Novel transformer-based self-supervised learning methods for improved HVAC fault diagnosis performance with limited labeled data. *Energy* 2023; **278**: 127972.

92 Ramirez-Loaiza ME, Sharma M, Kumar G, *et al.* Active learning: an empirical study of common baselines. *Data Min Knowl Disc* 2017; **31**: 287–313.

93 Zhang L, Wen J. Active learning strategy for high fidelity short-term data-driven building energy forecasting. *Energy Build* 2021; **244**: 111026.

94 Fan C, Wu Q, Zhao Y, *et al.* Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance. *Appl Energy* 2024; **356**: 122356.

95 Jacoby M, Tan SY, Henze G, *et al.* A high-fidelity residential building occupancy detection dataset. *Sci Data* 2021; **8**: 280.